北京工业大学学生开题报告表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 课题名称 | 金融行业用户画像与风险分析系统研究与实现 | | | | |
| 课题来源 | 企事业单位 | 课题类型 | BX | 导 师 | 付鹏斌 |
| 学生姓名 | 孟祥芾 | 学 号 | 20074210 | 专 业 | 计算机科学与技术 |
| 一、课题背景和意义  随着近年来越来越多贷款平台的开放以及银行贷款利率的下调，越来越多的人选择贷款消费，这会促进整体社会消费市场的流动，但对于放贷方来说，识别借款人还款的逾期风险就成了一个问题。信用不好的借款人，借了一大笔贷款金额，如果不能及时还款，那对于放贷方来说也会是一笔损失。因此，为了解决放贷方衡量借款人的金融风险等级，本文采用决策树的机器学习方法，学习高风险用户的自身特征，为放贷方制定一个系统对借款用户划分风险等级，并生成相关报告，从而实现辅助放贷方做出合理的金融决策。  二、课题研究现状  近年来，金融行业用户画像分析领域的研究主要体现在将用户画像应用在金融产品的推荐系统和营销上。例如，2017年，鲍忠铁在《现代商业银行》杂志上发表了题为《金融行业大数据用户画像实践》的论文，探讨了用户画像在金融企业中的应用和实践。2019年，姚鸣在对外经济贸易大学完成了博士学位论文，研究了基于大数据挖掘方法的互联网金融个人用户画像。他通过对某金融机构的业务核心网用户数据进行分析，并提出了基于用户画像的精准营销模型。同年，席文娅在南京农业大学完成了硕士学位论文，研究了基于用户画像的Z银行直销银行营销策略。她通过文献综述、数理统计和案例研究等方法，分析了某银行现有用户的基本属性和交易行为属性，构建了有效的用户画像，并提出了具体的营销策略。综上所述，这些研究为金融机构提供了更有效的营销手段和产品推广策略，推动了金融行业的发展和创新。  三、调研资料的准备  1.数据集，https://www.kaggle.com/datasets/ranadeep/credit-risk-dataset  2.类似实验经验https://blog.csdn.net/s09094031/article/details/92428209  四、设计目的  基于Java和Python开发一个金融行业风险分析系统。  五、设计要求  （1）对客户公开信息通过爬虫技术获取  （2）用神经网络进行客户分类  （3）基于大数据技术对同类客户进行风险分析  （4）给出客户风险等级和风险分析报告。系统开发严格按照软件工程的要求，进行需求分析、概要设计、系统和算法详细设计、系统开发和测试，最终撰写系统部署说明。  六、设计思路  1.数据集准备  阅读相关文献和任务书后，确定毕设的题目范围是“金融行业风险分析”，因此选择“贷款”相关的数据集。在找数据集的过程中首先想到从万方数据库中寻找，但尝试寻找一段时间后，发现该数据库中的数据大都是跟宏观的经济数据，很少有如此微观层面手机的用户数据。最后我选择从kaggle中寻找可用的公开数据集。  2.数据预处理  从kaggle中找到了一份名为“Credit Risk Analysis”的数据集，其中包括54700多名用户的近74个特性的贷款数据，如贷款金额、利率、还款期数、贷款等贷款信息；职称、工作年限、年收入、逾期次数等借款人信息；其中部分数据都为数值型数据，不适用与后面决策树机器模型 | | | | | |
| 使用，因此需要对数据集预处理，以便后续操作。同时该数据集中没有衡量用户是否为高风险用户的指标，所以我将用“当前还款金额除以贷款金额的值与0.4比较”用于衡量用户的风险等级，如果低于0.4，说明用户这笔贷款还的钱很少，可以认为他的风险等级比较高；如果高于0.4说明用户还了贷款的大部分金额，信用良好，风险等级较低。  3.决策树用户分类  得到一份可以使用的数据集后，我选择使用决策树作为本次毕业设计用到的机器学习模型，因为这个方法可以良好的适用于数据分类任务，通过学习数据的特征值将数据分类，很好地匹配毕设中“客户分类”的任务。在实际使用中，我计划将预处理中计算的得到的风险等级作为决策树模型中的标签值，选择其余数据列作为学习特征值，将数据集中70%可用的数据作为训练集，其余作为验证集。  4.分析报告生成系统  经过机器学习的分类后，我可以很容易根据用户的个人信息和贷款的相关信息得到用户的风险等级。有了这些信息后，我选择使用Freemarker等文档生成工具将上面这些数据进行可视化，展现给用户。我将编写一个报告模版，结合提到的生成工具，可以将风险等级、信贷信息、借款人信息等数据填写到文档中，从而实现自动化生成分析报告的功能。  5.结果验证  除了公开数据集外，我计划用python的selenium等工具包开发爬虫软件，对客户公开信息通过爬虫技术获取。获取数据后后，使用已开发的决策树模型分类、自动化生成报告，从而验证本次课程设计的成果。  七、系统实现  1.需求分析  需求来自于任务书中的要求，即要满足根据用户的特征描述，进行风险等级划分，并生成风险分析报告；在获取用户的特征数据时，除了可以使用公开数据集以外，也会用爬虫获取用户的公开信息。  2.概要设计  根据需求分析可知，该系统主要由三个部分组成：1爬虫获取用户相关信息，2根据用户基本信息进行画像分类，3根据风险等级自动生成报告。上述3个模块之间呈依赖关系，数据在系统中流通的过程中依次经过上述3个模块，因此在软件体系结构的设计上，我选择使用管道和过滤器体系结构，控制数据信息的流通方向，依次经过上述3个模块，实现整个系统的生命周期。  3.详细设计  a.爬虫获取用户信息  该模块实现的功能是根据用户的名字或手机号等比较具有标识性的信息，在网上的社交媒体（小红书、微博、知乎等）搜索该用户发布过的帖子，从帖子中提取有用的数据信息，比如收入、旅游次数、职业等。在后面的模块中，体现这些爬虫获取的信息，呈现给系统使用者。由上述模块功能，我设计该模块的实现过程为以下几个步骤：1社交平台关键词搜索用户，2识别用户发布内容，3发布内容中提取有价值信息。通过将上述3个步骤编码成一个固定的脚本，即可实现爬虫获取用户信息的功能需求。  b.风险等级分类  该模块的功能是实现一个机器学习模型，根据用户和贷款的相关信息，为用户划分风险等级。目前我已经从网络上找到一份合适数据集，其中包含分类任务需要用到的数据。我计划将从数据集中先只选择20个具有显著性的特征作为决策树节点划分的特征，因为过多的特征会使树的形状变得复杂，从而产生过拟合的问题。另外，在数据集的特征选择上，一些数值型数据需要提前预处理，按照值的标准划分成离散型数据，以便决策树的学习。必要时，也会引入随机森林、梯度提升决策树，以提高模型的稳定性和泛化能力。 | | | | | |
| c.自动化报告生成  该模块计划使用一个名为Freemarker的工具包，其实现的功能是自动化生成一个html代码的文档，文档中的部分关键字可以自动替换为指定的数据信息，生成的html语言文档，可以再通过外部工具生成word文档或pdf文档，从而实现生成风险分析报告的功能。具体设计为：1用word等可视化文档报告编辑工具，编写原始报告文档模版；2将文档模版中的关键信息用Freemarker语法替换；3用Java语言编写系统，将借款信息、用户信息、爬虫获取的信息、风险等级等需要填充到报告中的信息，汇总到一个Map中；4按Freemarker工具包的教程，将2步中的模版和3步得到的Map结合，生成最终需要的文档。  4.编码实现  系统计划主要使用python和Java两种编程语言实现，其中爬虫和机器学习部分使用python语言，自动化报告生成和系统的图形化界面使用Java。后期三部分各自都实现后，将决策树模型和爬虫脚本部署到Java项目中，实现三部分的统一结合。  5.测试  测试将主要分为单元测试和集成测试，其中单元测试分别测试详细设计中的3个模块是否都可以正常运行；集成测试用于验收最终的项目是否可以实现对目标用户先进行爬虫获取信息，再经过机器学习判断风险等级，最后自动化生成风险分析报告的全流程。  对于a模块的单元测试，输入数据为用户手机号或姓名，输出数据用户的其他信息，比如月收入、旅行次数等；对b模块的单元测试，输入数据为用户的20条特征描述，输出数据为用户的风险等级；c模块的单元测试，输入数据为a、b模块的输出数据和b模块的输入数据，输出数据为一个word或pdf风险等级分析报告。集成测试的输入数据为a、b模块的输入数据，输出为c模块的输出数据。  八、预期成果  完成设计要求，达到设计目的；按照要求完成相关文档的编写、实现要求功能的源代码、能运行的程序；完成毕业论文的撰写。  九、内容及时间安排  第一阶段：选题阶段（2023年12月-2024年1月）  完成对课题的调研，通过查阅文献等方式了解课题需求，同时对课题实验进行初步设计。  第二阶段：开题阶段（2024年1月-2024年2月）  在指导老师的帮助下顺利开题，完成对该课题的调研和设计，并完成开题报告及开通答辩。找到合适的数据集并进行预处理工作，为后续机器学习模型开发提供数据支持；熟悉开发环境并完成系统配置，为后续实际开发提供环境支持。  第三阶段：课题实现阶段（2024年2月-2024年4月）  进行题目算法的实现和设计，基本完成毕业设计所需的程序方面任务，同时在实验过程中继续学习相关内容并进行整理和汇总，完成论文初稿。  第四阶段：论文撰写及答辩阶段（2024年4月-2024年5月）  进行模型的改进及测试，即完成模型训练和数据集验证，并对结果进行理论分析，最终得出该实验的相关结论。在初稿的基础上完善毕业论文，形成定稿并准备答辩。  十、完成设计（论文）所具备的条件因素  指导老师的理论指导，查询相关文献，必备的硬件及软件环境，严格的时间安排，以及认真严谨的学习研究态度。  指导教师签名： 日期：2024年02月23日 | | | | | |