东莞农村商业银行客户生态管理系统项目

模型方案设计说明书

|  |  |  |  |
| --- | --- | --- | --- |
| 文档编号： |  |  |  |
| 编 撰： |  | 编撰日期： | 2023.6.2 |
| 审 核： |  | 审核日期： |  |
| 批 准： |  | 批准日期： |  |

目录

[1 简介 1](#_Toc516)

[1.1 项目背景 1](#_Toc8924)

[1.2 范围 1](#_Toc18041)

[1.3 术语定义 1](#_Toc2144)

[1.4 参考文档 1](#_Toc16098)

[2 业务目标 1](#_Toc24425)

[2.1 公司数据模型 1](#_Toc18333)

[2.1.1 流失客户预警 1](#_Toc12245)

[2.1.2 上下游潜客推荐 2](#_Toc23077)

[2.2 零售数据模型 2](#_Toc8701)

[2.2.1 交叉销售 2](#_Toc26382)

[2.2.2 商机挖掘 2](#_Toc12096)

[3 模型方案 2](#_Toc2287)

[3.1 公司数据模型 2](#_Toc15295)

[3.1.1 流失客户预警模型 2](#_Toc8101)

[3.1.1.1 数据模型目标 2](#_Toc2850)

[3.1.1.2 数据指标 3](#_Toc7819)

[3.1.1.3 数据处理 4](#_Toc20576)

[3.1.1.4 模型实现 5](#_Toc28063)

[3.1.1.5 模型验证 6](#_Toc22746)

[3.1.1.6 模型部署 6](#_Toc31919)

[3.1.2 上下游潜客推荐 6](#_Toc5457)

[3.1.2.1 数据模型目标 6](#_Toc24828)

[3.1.2.2 数据指标 6](#_Toc6007)

[3.1.2.3 数据处理 7](#_Toc22782)

[3.1.2.4 模型实现 7](#_Toc4403)

[3.1.2.5 模型验证 9](#_Toc12070)

[3.1.2.6 模型部署 9](#_Toc2763)

[3.2 零售数据模型 9](#_Toc31697)

[3.2.1 交叉销售模型 9](#_Toc28559)

[3.2.1.1 数据模型目标 9](#_Toc2003)

[3.2.1.2 数据指标 9](#_Toc19648)

[3.2.1.3 数据处理 12](#_Toc23607)

[3.2.1.4 模型实现 13](#_Toc15559)

[3.2.1.5 模型上线验证 14](#_Toc12588)

[3.2.1.6 模型部署 14](#_Toc2288)

[3.2.2 商机挖掘模型 15](#_Toc15675)

[3.2.2.1 数据模型目标 15](#_Toc7756)

[3.2.2.2 数据指标 15](#_Toc12079)

[3.2.2.3 数据处理 18](#_Toc5743)

[3.2.2.4 模型实现 19](#_Toc12389)

[3.2.2.5 模型上线验证 20](#_Toc23314)

[3.2.2.6 模型部署 20](#_Toc31314)

[4 模型部署环境要求 20](#_Toc25579)

[5 模型实施进度安排 21](#_Toc8579)

# 

# 简介

## 项目背景

为贯彻落实“1+12348”战略规划，推动科技创新及公私业务数字化转型，我行计划打造从“客户筛选-客户数据分析-产业链条关联-产品精准推送-营销跟踪管理”的全流程、全链条、智能化的客户经营生态系统（下面称“新CRM系统”），全方位打通公私业务壁垒，丰富营销业务场景、提升客户关系管理和数据分析，加强客户经营管理能力。

新CRM系统建设将以旧CRM系统业务功能为基础，遵循集约化、整体化建设思路，整合旧系统已有功能，借鉴同业先进经验，将新CRM系统建设成为全行客户的营销中心、数据中心、产能中心，并确保系统的先进性和可扩展性，满足未来3-5年的业务发展需求。

新CRM系统建设目标包括实现OCRM（操作型CRM）和ACRM（分析型CRM）的整合，提供客户效益分析和预测、交叉营销、产品及服务使用分析、客户流失分析等，为数据营销提供决策中心。依托新CRM提供的数据指标，分析客户的不同行为特征，建立面向不同业务场景的数据模型，通过数据支撑战略决策，推动业务发展。

## 范围

本说明书描述数据模型设计方案，包括业务目标、模型方案、模型部署环境要求等。

本说明书主要阅读对象是科技开发人员、建模人员等，需要综合考虑可实现性、易维护性等，进行综合考虑做设计和开发。

## 术语定义

无

## 参考文档

《东莞农村商业银行客户生态管理系统需求说明书》

# 业务目标

## 零售数据模型

### 流失客户预警

根据存量客户的基础属性、客户价值、产品持有、交易行为分析等信息，通过对客户信息进行特征分析，筛选与流失密切相关的特征，预测当前客户流失的概率，输出流失客户名单。

流失客户定义：客户当期金融资产月日均较近三月下降比例超过50%。50%为变量，可支持调节。

### 上下游潜客推荐

镇街产业客户名单为手工初始化名单，作为上下游潜客推荐模型的输入名单，通过将镇街产业客户名单与存量客户结合（包括镇街产业客户与行内存量客户的交易特征，存量客户与行外客户的交易特征），分析客户基础信息、交易特征等信息，建立数据模型，输出潜客名单。

潜客定义：月均交易笔数大于3笔，月均交易金额大于300万元。月均交易笔数和月均交易金额可支持调节。

## 零售数据模型

### 交叉销售

产品交叉销售是给持有A产品的客户推荐B产品或者其他产品，通过对客户的基础信息、交易行为信息、产品信息等进行特征分析，选取最优产品组合，推荐给客户，输出客户以及对应的推荐产品。

### 商机挖掘

从八大类产品中选择存款，保险产品，进行商机挖掘推荐，通过对客户基础信息、历史购买信息等进行客户特征分析，预测当前未购买该产品的客户未来购买的可能性，从而进行产品推荐，输出目标客户名单。

# 模型方案

## 公司数据模型

### 流失客户预警模型

#### 数据模型目标

通过分析存量客户的基础属性、交易行为等信息，建立数据模型，预测未来可能要流失的客户。

#### 数据字段

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 字段 | 序号 | 字段 |
| 1 | 客户编号 | 36 | 票据池签约日期 |
| 2 | 客户名称 | 37 | 秒贴签约日期 |
| 3 | 员工人数 | 38 | 代缴社保签约日期 |
| 4 | 企业规模 | 39 | 代缴公积金签约日期 |
| 5 | 企业性质 | 40 | 公积金缴存笔数 |
| 6 | 开户年限 | 41 | 公积金缴存金额 |
| 7 | 企业名称变更次数 | 42 | 企业手机银行交易笔数 |
| 8 | 经营范围变更次数 | 43 | 企业手机银行交易金额 |
| 9 | 经营地址变更次数 | 44 | 现金管理平台交易笔数 |
| 10 | 本月日均存款余额 | 45 | 现金管理平台交易金额 |
| 11 | 活期存款月日均 | 46 | 园区通交易笔数 |
| 12 | 通知存款月日均 | 47 | 园区通交易金额 |
| 13 | 协议存款月日均 | 48 | 银校通交易笔数 |
| 14 | 保证金存款月日均 | 49 | 银校通交易金额 |
| 15 | “定享通”定期存款月日均 | 50 | 教培系统交易笔数 |
| 16 | 大额存单月日均 | 51 | 教培系统交易金额 |
| 17 | 结构性存款月日均 | 52 | 单位结算卡笔数 |
| 18 | 本月日均贷款余额 | 53 | 单位结算卡金额 |
| 19 | 本月日均保本理财余额 | 54 | POS收单笔数 |
| 20 | 本月日均非保本理财余额 | 55 | POS收单金额 |
| 21 | 本月日均贴现余额 | 56 | 对公ETC交易次数 |
| 22 | 本月日均开票余额 | 57 | 对公ETC交易金额 |
| 23 | 工资代发签约日期 | 58 | 对公银银转账交易次数 |
| 24 | 公积金缴存签约日期 | 59 | 对公银银转账交易金额 |
| 25 | 企业网银签约日期 | 60 | 代缴社保笔数 |
| 26 | 企业手机银行签约日期 | 61 | 代缴社保金额 |
| 27 | 现金管理平台签约日期 | 62 | 代缴公积金笔数 |
| 28 | 园区通签约日期 | 63 | 代缴公积金金额 |
| 29 | 银校通签约日期 | 64 | 企业网银交易笔数 |
| 30 | 产融平台签约日期 | 65 | 企业网银交易金额 |
| 31 | 教培系统签约日期 | 66 | 柜面交易笔数 |
| 32 | 单位结算卡签约日期 | 67 | 柜面交易金额 |
| 33 | POS收单开通日期 | 68 | 金融资产月日均 |
| 34 | 开通财票通 | 69 | 近三个月金融资产月日均 |
| 35 | 投标通签约日期 |  |  |

#### 数据处理

数据处理过程包括数据准备和数据处理步骤。

##### 数据准备

真实数据要求有10000条以上用于模型训练。

现有数据来源广泛且格式复杂，获取数据的初期需要根据业务需求，针对客户提供的数据，完成数据预处理、转换、清理等工作，抽取数据特征，在全量数据中抽取适用于特定模型的数据维度，分析数据分布情况、数据类型以及数据质量，从而更好的支持模型的创建。

##### 数据处理步骤

1. **空值处理**

实际数据库中，属性值缺失的情况经常发全甚至是不可避免的。因此，在大多数情况下，数据集是不完整的，因此需要在进行数据分析数据准备阶段对空值（缺失值）进行处理。

1. **去重**

数据集中相同数据记录重复出现时保留一条记录。

1. **类型转换**

将字段的数据类型转换为另一种类型，如字符串转为数值型。

1. **删除列**

对选中的列进行删除操作。

1. **标签列确定**

结合流失客户定义，基于金融资产月日均以及近三个月金融资产月日均两个字段生成标签列，用于模型训练与模型评估。

#### 模型实现

##### 模型建立

在建模阶段，需要结合实际场景构建各类数据挖掘模型，以对数据和信息进行梳理、简化、抽取和归纳，从而得到效果较为理想的合理化模型。

##### 构建算法模型

该模型目标是预测潜在流失客户，最终输出0或1表示流失或者不流失，因此采用二分类模型。

常用的分类预测模型如下：

* **逻辑回归**

逻辑回归又称logistic[回归分析](https://baike.baidu.com/item/%E5%9B%9E%E5%BD%92%E5%88%86%E6%9E%90)，是一种广义的线性回归分析模型，常用于数据挖掘，疾病自动诊断，经济预测等领域。例如，探讨引发疾病的危险因素，并根据危险因素预测疾病发生的概率等。以胃癌病情分析为例，选择两组人群，一组是胃癌组，一组是非胃癌组，两组人群必定具有不同的体征与生活方式等。因此[因变量](https://baike.baidu.com/item/%E5%9B%A0%E5%8F%98%E9%87%8F)就为是否胃癌，值为“是”或“否”，自变量就可以包括很多了，如年龄、性别、饮食习惯、[幽门螺杆菌](https://baike.baidu.com/item/%E5%B9%BD%E9%97%A8%E8%9E%BA%E6%9D%86%E8%8F%8C)感染等。自变量既可以是连续的，也可以是分类的。然后通过logistic回归分析，可以得到自变量的权重，从而可以大致了解到底哪些因素是胃癌的危险因素。同时根据该权值可以根据危险因素预测一个人患癌症的可能性。

* **SVM**

一类按监督学习方式对数据进行二元分类的广义线性分类器,其决策边界是对学习样本求解的最大边距超平面。

* **随机森林**

随机森林可以应用在分类和回归问题上。实现这一点，取决于随机森林的每颗cart树是分类树还是回归树，如果cart树是分类数，那么采用的计算原则就是gini指数。随机森林基于每棵树的分类结果，采用多数表决的手段进行分类。

* **XGBoost**

Boosting算法的思想是将许多弱分类器集成在一起形成一个强分类器。因为XGBoost是一种提升树模型，所以它是将许多树模型集成在一起，形成一个很强的分类器。

##### 3.1.1.4.3模型搭建与训练

选取一种分类模型，训练模型并预测未来业务需求指标，输出客户流失预警信息。

训练过程中调参优化模型，通过查看模型正确率，确定模型的最优版本。

#### 模型验证

我们采用准确率作为模型上线后验证的指标。因为训练数据是历史数据，验证时应使用近期最新的数据验证模型，这样可以体现模型对未来预测的能力。具体地，利用流失客户定义得到每个客户是否为流失客户的标签。利用我们的模型对数据进行预测，得到每个用户是否为流失客户的预测结果，该结果和上面的真实标签进行对比，可以验证模型的准确程度。

#### 模型部署与使用

模型以Python程序的形式进行部署，在使用时可运行该程序，程序的输入是.del格式的数据文件，输出是.txt格式的结果文件。

### 上下游潜客推荐

#### 数据模型目标

通过分析客户交易数据，注册地址，关联关系，建立分类模型，使潜客名单能够分类，推出潜客名单。

#### 数据字段

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 字段 | 序号 | 字段 |
| 1 | 客户编号 | 8 | 近3个⽉交易笔数 |
| 2 | 近1个⽉交易⾦额 | 9 | 近1年交易笔数 |
| 3 | 近3个⽉交易⾦额 | 10 | ⽉均交易笔数 |
| 4 | 近1年交易⾦额 | 11 | 注册地址 |
| 5 | ⽉均交易⾦额 | 12 | 经营地址 |
| 6 | 交易笔数 | 13 | 关联关系 |
| 7 | 近1个月交易笔数 |  |  |

#### 数据处理

数据处理过程包括数据准备和数据处理。

##### 数据准备

真实数据要求有10000条以上用于模型训练。

现有数据来源广泛且格式复杂，获取数据的初期需要根据业务需求，针对客户提供的数据，完成数据预处理、转换、清理等工作，抽取数据特征，在全量数据中抽取适用于特定模型的数据维度，分析数据分布情况、数据类型以及数据质量，从而更好的支持模型的创建。

##### 数据处理步骤

1. **空值处理**

实际数据库中，属性值缺失的情况经常发全甚至是不可避免的。因此，在大多数情况下，数据集是不完整的，因此需要在进行数据分析数据准备阶段对空值（缺失值）进行处理。

1. **去重**

数据集中相同数据记录重复出现时保留一条记录。

1. **类型转换**

将字段的数据类型转换为另一种类型，如字符串转为数值型。

1. **删除列**

对选中的列进行删除操作。

1. **构建训练集标签**

利用潜客的定义可以得到每个客户是否为潜客的标签。

#### 模型实现

##### 模型建立

潜客推荐算法流程分为以下几步：

1. 数据预处理
2. 分类模型训练
3. 根据模型给与潜客分类

##### 构建算法模型

由于该模型需求是潜客推荐模型，所以采用二分类模型，输出是潜在客户或者不是潜在客户，常用的分类预测模型如下：

* **逻辑回归**

逻辑回归又称logistic[回归分析](https://baike.baidu.com/item/%E5%9B%9E%E5%BD%92%E5%88%86%E6%9E%90)，是一种广义的线性回归分析模型，常用于数据挖掘，疾病自动诊断，经济预测等领域。例如，探讨引发疾病的危险因素，并根据危险因素预测疾病发生的概率等。以胃癌病情分析为例，选择两组人群，一组是胃癌组，一组是非胃癌组，两组人群必定具有不同的体征与生活方式等。因此[因变量](https://baike.baidu.com/item/%E5%9B%A0%E5%8F%98%E9%87%8F)就为是否胃癌，值为“是”或“否”，自变量就可以包括很多了，如年龄、性别、饮食习惯、[幽门螺杆菌](https://baike.baidu.com/item/%E5%B9%BD%E9%97%A8%E8%9E%BA%E6%9D%86%E8%8F%8C)感染等。自变量既可以是连续的，也可以是分类的。然后通过logistic回归分析，可以得到自变量的权重，从而可以大致了解到底哪些因素是胃癌的危险因素。同时根据该权值可以根据危险因素预测一个人患癌症的可能性。

* **SVM**

一类按监督学习方式对数据进行二元分类的广义线性分类器,其决策边界是对学习样本求解的最大边距超平面。

* **随机森林**

随机森林可以应用在分类和回归问题上。实现这一点，取决于随机森林的每颗cart树是分类树还是回归树，如果cart树是分类数，那么采用的计算原则就是gini指数。随机森林基于每棵树的分类结果，采用多数表决的手段进行分类。

* **XGBoost**

Boosting算法的思想是将许多弱分类器集成在一起形成一个强分类器。因为XGBoost是一种提升树模型，所以它是将许多树模型集成在一起，形成一个很强的分类器。

##### 模型搭建与训练

该模型选取线性回归模型或者逻辑回归模型，训练模型并预测未来业务需求指标，输出客户流失预警信息。

训练过程中调参优化模型，通过查看模型正确率，确定模型的最优版本。

#### 模型验证

我们采用准确率作为模型上线后验证的指标。因为训练数据是历史数据，验证时应使用近期最新的数据验证模型，这样可以体现模型对未来预测的能力。根据定义月均交易笔数大于3笔，月均交易金额大于300万元，可以得到每个用户是否为潜客的标签，然后利用模型对这些数据进行预测，得到的结果和上面的标签进行对比，可以得到模型准确性。

#### 模型部署与使用

模型以Python程序的形式部署在用户提供的机器上，用户使用时运行该程序，程序的输入是.del格式的数据文件，输出是.txt格式的结果文件。

### 交叉销售模型

#### 数据模型目标

通过分析客户的基础信息、交易行为等信息，建立数据模型，能够选取最优产品组合，便于为客户推荐对应产品。

#### 数据字段

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 字段 | 序号 | 字段 |
| 1 | 客户编号 | 68 | 贷款月日均 |
| 2 | 客户名称 | 69 | 贷款季日均 |
| 3 | 年龄 | 70 | 历史贷款余额 |
| 4 | 性别 | 71 | 消费贷款余额 |
| 5 | 开户年限 | 72 | 消费贷款月日均 |
| 6 | 学历 | 73 | 消费贷款季日均 |
| 7 | 婚姻状况 | 74 | 历史消费贷款余额 |
| 8 | 村民 | 75 | 按揭贷款余额 |
| 9 | 新市民 | 76 | 按揭贷款月日均 |
| 10 | 代发客户 | 77 | 按揭贷款季日均 |
| 11 | 按揭客户 | 78 | 历史按揭贷款余额 |
| 12 | 大学生客户 | 79 | 个人经营性贷款余额 |
| 13 | 股东客户 | 80 | 个人经营性贷款月日均 |
| 14 | 企业主客户 | 81 | 个人经营性贷款季日均 |
| 15 | AUM余额 | 82 | 历史个人经营性贷款余额 |
| 16 | AUM月日均 | 83 | 网络贷款余额 |
| 17 | AUM季日均 | 84 | 网络贷款月日均 |
| 18 | 历史AUM余额 | 85 | 网络贷款季日均 |
| 19 | 存款余额 | 86 | 历史网络贷款余额 |
| 20 | 存款月日均 | 87 | 信用卡贷款月日均 |
| 21 | 存款季日均 | 88 | 信用卡分期余额 |
| 22 | 历史存款余额 | 89 | 信用卡分期月日均 |
| 23 | 活期存款余额 | 90 | 信用卡分期季日均 |
| 24 | 活期存款月日均 | 91 | 历史信用卡分期余额 |
| 25 | 活期存款季日均 | 92 | 持有金融产品个数 |
| 26 | 历史活期存款余额 | 93 | 持有渠道产品个数 |
| 27 | 普通定期余额 | 94 | 签约消费渠道个数 |
| 28 | 普通定期月日均 | 95 | 持有定期存款产品个数 |
| 29 | 普通定期季日均 | 96 | 持有贷款产品个数 |
| 30 | 历史普通定期余额 | 97 | 持有理财产品个数 |
| 31 | 大额存单余额 | 98 | 持有保险产品个数 |
| 32 | 大额存单月日均 | 99 | 持有基金产品个数 |
| 33 | 大额存单季日均 | 100 | 持有国债产品个数 |
| 34 | 历史大额存单余额 | 101 | 持有高净值产品个数 |
| 35 | 安心存单余额 | 102 | 手机银行近三个月登陆次数 |
| 36 | 安心存单月日均 | 103 | 客户90天交易频率 |
| 37 | 安心存单季日均 | 104 | 客户180天交易频率 |
| 38 | 历史安心存单余额 | 105 | 近三月转入金额 |
| 39 | 理财余额 | 106 | 近三月转入笔数 |
| 40 | 理财月日均 | 107 | 近三月转出金额 |
| 41 | 理财季日均 | 108 | 近三月转出笔数 |
| 42 | 历史理财余额 | 109 | 近三月基金买入金额（不含定投） |
| 43 | 保险余额 | 110 | 近三月基金买入笔数（不含定投） |
| 44 | 保险月日均 | 111 | 近三月基金赎回金额 |
| 45 | 保险季日均 | 112 | 近三月基金赎回笔数 |
| 46 | 历史保险余额 | 113 | 本月网银跨行转账转出金额 |
| 47 | 基金余额 | 114 | 本月手机银行跨行转账转出金额 |
| 48 | 基金月日均 | 115 | 柜台近3月交易笔数 |
| 49 | 基金季日均 | 116 | 手机银行近3月交易笔数 |
| 50 | 历史基金余额 | 117 | 微信近三月转入笔数 |
| 51 | 积利宝余额 | 118 | 微信近三月转入金额 |
| 52 | 积利宝月日均 | 119 | 微信近三月转出笔数 |
| 53 | 积利宝季日均 | 120 | 微信近三月转出金额 |
| 54 | 历史积利宝余额 | 121 | 支付宝近三月转入笔数 |
| 55 | 积存金余额 | 122 | 支付宝近三月转入金额 |
| 56 | 积存金月日均 | 123 | 支付宝近三月转出笔数 |
| 57 | 积存金季日均 | 124 | 支付宝近三月转出金额 |
| 58 | 历史积存金余额 | 125 | POS近三月转入笔数 |
| 59 | 国债余额 | 126 | POS近三月转入金额 |
| 60 | 国债月日均 | 127 | 银银转账近三月转入金额 |
| 61 | 国债季日均 | 128 | 银银转账近三月转出金额 |
| 62 | 历史国债余额 | 129 | 风险评级 |
| 63 | 高净值产品余额 | 130 | 储蓄产品偏好 |
| 64 | 高净值产品月日均 | 131 | 理财产品偏好 |
| 65 | 高净值产品季日均 | 132 | 保险产品偏好 |
| 66 | 历史高净值产品余额 | 133 | 基金产品偏好 |
| 67 | 贷款余额 | 134 | 交易渠道偏好 |

#### 数据处理

数据处理过程包括数据准备和数据处理步骤

##### 数据准备

真实数据提供的时间较晚，项目开发是在假数据上面进行的，所以当提供真实数据时，应保持与假数据的数据字段一致。真实数据要求有10000条以上用于模型训练。

现有数据来源广泛且格式复杂，获取数据的初期需要根据业务需求，针对客户提供的数据，完成数据预处理、转换、清理等工作，抽取数据特征，在全量数据中抽取适用于特定模型的数据维度，分析数据分布情况、数据类型以及数据质量，从而更好的支持模型的创建。

##### 数据处理步骤

1. **空值处理**

实际数据库中，属性值缺失的情况经常发全甚至是不可避免的。因此，在大多数情况下，数据集是不完整的，因此需要在进行数据分析数据准备阶段对空值（缺失值）进行处理。

1. **去重**

数据集中相同数据记录重复出现时保留一条记录。

1. **类型转换**

将字段的数据类型转换为另一种类型，如字符串转为数值型。

1. **删除列**

对选中的列进行删除操作。

#### 模型实现

##### 模型建立

推荐算法流程分为两个阶段：训练阶段和推荐阶段。

对于训练阶段，可分为以下几步：

1. 数据预处理，建立user-item表
2. 建立商品整体共现矩阵
3. 建立物品相似度矩阵

对于推荐阶段，可分为以下几步：

1. 寻找与被推荐用户喜爱物品集最相似的N个物品
2. 计算用户对这N个物品感兴趣程序列表并逆序排列

##### 构建算法模型

由于该模型需求是零售数据模型，所以采用推荐算法模型，常用的推荐模型如下：

* **协同过滤**

协同过滤（collaborative filtering）是一种在推荐系统中广泛使用的技术。该技术通过分析用户或者事物之间的相似性，来预测用户可能感兴趣的内容并将此内容推荐给用户。这里的相似性可以是人口特征的相似性，也可以是历史浏览内容的相似性，还可以是个人通过一定机制给与某个事物的回应。比如，A和B是无话不谈的好朋友，并且都喜欢看电影，那么协同过滤会认为A和B的相似度很高，会将A喜欢但是B没有关注的电影推荐给B，反之亦然。

协同过滤推荐分为3种类型：

* 基于用户(user-based)的协同过滤(UserCF)
* 基于物品(item-based)的协同过滤（ItemCF算法)
* 基于模型(model-based)的协同过滤 (ModelCF算法)
* **关联规则**

关联分析又称关联挖掘，就是在交易数据、关系数据或其他信息载体中，查找存在于项目集合或对象集合之间的频繁模式、关联、相关性或因果结构。

关联分析是一种简单、实用的分析技术，就是发现存在于大量数据集中的关联性或相关性，从而描述了一个事物中某些属性同时出现的规律和模式。

关联分析是从大量数据中发现项集之间有趣的关联和相关联系。

* **Apriori算法**

电子商务中常用的一种数据挖掘方法就是从用户交易数据集中寻找商品之间的关联规则。关联规则中常用的一种算法是Apriori算法。该算法主要包含两个步骤：首先找出数据集中所有的频繁项集，这些项集出现的频繁性要大于或等于最小支持度；然后根据频繁项集产生强关联规则，这些规则必须满足最小支持度和最小置信度。

##### 模型搭建与训练

该模型选取推荐模型，训练模型并推出客户名单和对应推荐产品。

训练过程中调参优化模型，通过查看模型正确率，确定模型的最优版本。

#### 模型上线验证

由于该模型用于辅助商品营销的用途，模型的效果不容易用客观的评价指标来衡量，使用者可以使用该模型的结果(即向客户推荐的产品)来实践验证向该客户推荐这些产品是否符合客户需求。

#### 模型部署与使用

模型以Python程序的形式部署在用户提供的机器上，用户使用时运行该程序，程序的输入是.del格式的数据文件，输出是.txt格式的结果文件。

### 商机挖掘模型

#### 数据模型目标

通过分析历史购买存款和保险产品的客户基础信息、购买信息，建立数据模型，能够预测潜客名单。

#### 数据字段

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 字段 | 序号 | 字段 |
| 1 | 客户编号 | 68 | 贷款月日均 |
| 2 | 客户名称 | 69 | 贷款季日均 |
| 3 | 年龄 | 70 | 历史贷款余额 |
| 4 | 性别 | 71 | 消费贷款余额 |
| 5 | 开户年限 | 72 | 消费贷款月日均 |
| 6 | 学历 | 73 | 消费贷款季日均 |
| 7 | 婚姻状况 | 74 | 历史消费贷款余额 |
| 8 | 村民 | 75 | 按揭贷款余额 |
| 9 | 新市民 | 76 | 按揭贷款月日均 |
| 10 | 代发客户 | 77 | 按揭贷款季日均 |
| 11 | 按揭客户 | 78 | 历史按揭贷款余额 |
| 12 | 大学生客户 | 79 | 个人经营性贷款余额 |
| 13 | 股东客户 | 80 | 个人经营性贷款月日均 |
| 14 | 企业主客户 | 81 | 个人经营性贷款季日均 |
| 15 | AUM余额 | 82 | 历史个人经营性贷款余额 |
| 16 | AUM月日均 | 83 | 网络贷款余额 |
| 17 | AUM季日均 | 84 | 网络贷款月日均 |
| 18 | 历史AUM余额 | 85 | 网络贷款季日均 |
| 19 | 存款余额 | 86 | 历史网络贷款余额 |
| 20 | 存款月日均 | 87 | 信用卡贷款月日均 |
| 21 | 存款季日均 | 88 | 信用卡分期余额 |
| 22 | 历史存款余额 | 89 | 信用卡分期月日均 |
| 23 | 活期存款余额 | 90 | 信用卡分期季日均 |
| 24 | 活期存款月日均 | 91 | 历史信用卡分期余额 |
| 25 | 活期存款季日均 | 92 | 持有金融产品个数 |
| 26 | 历史活期存款余额 | 93 | 持有渠道产品个数 |
| 27 | 普通定期余额 | 94 | 签约消费渠道个数 |
| 28 | 普通定期月日均 | 95 | 持有定期存款产品个数 |
| 29 | 普通定期季日均 | 96 | 持有贷款产品个数 |
| 30 | 历史普通定期余额 | 97 | 持有理财产品个数 |
| 31 | 大额存单余额 | 98 | 持有保险产品个数 |
| 32 | 大额存单月日均 | 99 | 持有基金产品个数 |
| 33 | 大额存单季日均 | 100 | 持有国债产品个数 |
| 34 | 历史大额存单余额 | 101 | 持有高净值产品个数 |
| 35 | 安心存单余额 | 102 | 手机银行近三个月登陆次数 |
| 36 | 安心存单月日均 | 103 | 客户90天交易频率 |
| 37 | 安心存单季日均 | 104 | 客户180天交易频率 |
| 38 | 历史安心存单余额 | 105 | 近三月转入金额 |
| 39 | 理财余额 | 106 | 近三月转入笔数 |
| 40 | 理财月日均 | 107 | 近三月转出金额 |
| 41 | 理财季日均 | 108 | 近三月转出笔数 |
| 42 | 历史理财余额 | 109 | 近三月基金买入金额（不含定投） |
| 43 | 保险余额 | 110 | 近三月基金买入笔数（不含定投） |
| 44 | 保险月日均 | 111 | 近三月基金赎回金额 |
| 45 | 保险季日均 | 112 | 近三月基金赎回笔数 |
| 46 | 历史保险余额 | 113 | 本月网银跨行转账转出金额 |
| 47 | 基金余额 | 114 | 本月手机银行跨行转账转出金额 |
| 48 | 基金月日均 | 115 | 柜台近3月交易笔数 |
| 49 | 基金季日均 | 116 | 手机银行近3月交易笔数 |
| 50 | 历史基金余额 | 117 | 微信近三月转入笔数 |
| 51 | 积利宝余额 | 118 | 微信近三月转入金额 |
| 52 | 积利宝月日均 | 119 | 微信近三月转出笔数 |
| 53 | 积利宝季日均 | 120 | 微信近三月转出金额 |
| 54 | 历史积利宝余额 | 121 | 支付宝近三月转入笔数 |
| 55 | 积存金余额 | 122 | 支付宝近三月转入金额 |
| 56 | 积存金月日均 | 123 | 支付宝近三月转出笔数 |
| 57 | 积存金季日均 | 124 | 支付宝近三月转出金额 |
| 58 | 历史积存金余额 | 125 | POS近三月转入笔数 |
| 59 | 国债余额 | 126 | POS近三月转入金额 |
| 60 | 国债月日均 | 127 | 银银转账近三月转入金额 |
| 61 | 国债季日均 | 128 | 银银转账近三月转出金额 |
| 62 | 历史国债余额 | 129 | 风险评级 |
| 63 | 高净值产品余额 | 130 | 储蓄产品偏好 |
| 64 | 高净值产品月日均 | 131 | 理财产品偏好 |
| 65 | 高净值产品季日均 | 132 | 保险产品偏好 |
| 66 | 历史高净值产品余额 | 133 | 基金产品偏好 |
| 67 | 贷款余额 | 134 | 交易渠道偏好 |

#### 数据处理

数据处理过程包括数据准备和数据处理步骤。

##### 数据准备

真实数据要求有10000条以上用于模型训练。

现有数据来源广泛且格式复杂，获取数据的初期需要根据业务需求，针对客户提供的数据，完成数据预处理、转换、清理等工作，抽取数据特征，在全量数据中抽取适用于特定模型的数据维度，分析数据分布情况、数据类型以及数据质量，从而更好的支持模型的创建。

##### 数据处理步骤

1. **空值处理**

实际数据库中，属性值缺失的情况经常发全甚至是不可避免的。因此，在大多数情况下，数据集是不完整的，因此需要在进行数据分析数据准备阶段对空值（缺失值）进行处理。

1. **去重**

数据集中相同数据记录重复出现时保留一条记录。

1. **类型转换**

将字段的数据类型转换为另一种类型，如字符串转为数值型。

1. **删除列**

对选中的列进行删除操作。

#### 模型实现

##### 模型建立

推荐算法流程分为两个阶段：训练阶段和推荐阶段。

对于训练阶段，可分为以下几步：

1. 数据预处理，建立user-item表
2. 建立商品整体共现矩阵
3. 建立物品相似度矩阵

对于推荐阶段，可分为以下几步：

1. 寻找与被推荐用户喜爱物品集最相似的N个物品
2. 计算用户对这N个物品感兴趣程序列表并逆序排列

##### 构建算法模型

由于该模型需求是零售数据模型，所以采用推荐算法模型，常用的推荐模型如下：

* **协同过滤**

协同过滤（collaborative filtering）是一种在推荐系统中广泛使用的技术。该技术通过分析用户或者事物之间的相似性，来预测用户可能感兴趣的内容并将此内容推荐给用户。这里的相似性可以是人口特征的相似性，也可以是历史浏览内容的相似性，还可以是个人通过一定机制给与某个事物的回应。比如，A和B是无话不谈的好朋友，并且都喜欢看电影，那么协同过滤会认为A和B的相似度很高，会将A喜欢但是B没有关注的电影推荐给B，反之亦然。

协同过滤推荐分为3种类型：

* 基于用户(user-based)的协同过滤(UserCF)
* 基于物品(item-based)的协同过滤（ItemCF算法)
* 基于模型(model-based)的协同过滤 (ModelCF算法)
* **关联规则**

关联分析又称关联挖掘，就是在交易数据、关系数据或其他信息载体中，查找存在于项目集合或对象集合之间的频繁模式、关联、相关性或因果结构。

关联分析是一种简单、实用的分析技术，就是发现存在于大量数据集中的关联性或相关性，从而描述了一个事物中某些属性同时出现的规律和模式。

关联分析是从大量数据中发现项集之间有趣的关联和相关联系。

* **Apriori算法**

电子商务中常用的一种数据挖掘方法就是从用户交易数据集中寻找商品之间的关联规则。关联规则中常用的一种算法是Apriori算法。该算法主要包含两个步骤：首先找出数据集中所有的频繁项集，这些项集出现的频繁性要大于或等于最小支持度；然后根据频繁项集产生强关联规则，这些规则必须满足最小支持度和最小置信度。

##### 模型搭建与训练

该模型选取推荐模型，训练模型并推出客户名单和对应推荐产品。

训练过程中调参优化模型，通过查看模型正确率，确定模型的最优版本。

#### 模型上线验证

该模型的目的是找到某产品的目标客户，和上面的交叉销售逻辑一致，使用者可以利用模型结果指示的客户名单，向这些客户营销该产品，如果符合客户的需求，那么证明该模型是有效的。

#### 模型部署与使用

模型以Python程序的形式部署在用户提供的机器上，用户使用时运行该程序，程序的输入是.del格式的数据文件，输出是.txt格式的结果文件。

# 模型部署环境要求

（1）硬件环境：

CPU：Intel Core i7 10代以上

内存：32G以上

硬盘空间：100G以上

网络带宽：3Mbytes/s以上

1. 软件环境：

操作系统：Linux