

泰迪科技数据分析项目实训

广电大数据用户画像描绘（spark 实现）

培训解决方案

广东泰迪智能科技股份有限公司 版权所有

地址：广州市经济技术开发区开泰大道 36 号 1 栋 212

网址：<http://www.tipdm.com>

邮箱：services@tipdm.com

邮编：510663

联系人：

电话：

目录

1	项目介绍	3
1.1	项目背景.....	3
1.2	项目目标.....	4
1.3	项目数据.....	4
1.4	项目周期.....	4
1.5	项目难度.....	4
2	项目任务	5
3	项目流程	5
4	项目核心	5
5	实现工具	5
6	实训对象	6
7	前置知识	6
8	实训对应的就业岗位.....	6
8.1	就业岗位.....	6
8.2	岗位分析.....	6
9	实训收益	6
10	证书.....	6
11	附件一 前置课程课表	8
12	作业提交时间表	18
13	评分标准	20
	工作室邀请函	22

1 项目介绍

1.1 项目背景

随着互联网技术的快速发展和应用扩展，国家正式推进三网融合政策，三网融合是指电信网、广播电视网、互联网在向宽带通信网、数字电视网、下一代互联网演进过程中，三大网络通过技术改造，其技术功能趋于一致，业务范围趋于相同，网络互联互通、资源共享，能为用户提供语音、数据和广播电视等多种服务。随着三网融合的深入推进，IPTV（交互式网络电视）加速布局，OTT（指通过互联网向用户提供各种应用服务）风起云涌，新媒体业务的飞速发展对传统媒体造成了巨大冲击，广电行业依靠资源稀缺形成的垄断优势已经失去。

复杂激烈的竞争环境，使得广电的客户流失问题变得异常突出。如何减少客户流失、挽留客户并挖掘客户潜在需求，是广电公司目前急需解决的问题！

以往传统媒体播送时代，广电公司“不知道用户在哪里，不知道用户是谁，也不知道用户想看什么”，因此难以精准把握用户需求，而随着数字电视机顶盒等技术的普及，广电公司具备了获取用户身份信息数据、实时收视数据的能力。

现如今，广电公司已经积累了海量的用户数据，包括用户基本信息数据、用户收视数据、用户订单数据、用户账单数据等。因此广电公司可以根据用户的特点，从人群、时间、地点、产品和付费方式五个维度来挖掘分析用户数据，对用户进行全面的画像，比如从人群维度分析明确用户的年龄特征，如少儿、青少年、中年或老年等，以及分析收视语言是外语、国语、粤语等；从时间维度分析用户每天观看电视的时长或者用户观看某一电视节目的时长；从地点维度分析明确用户的收视常在地；从产品维度分析用户喜欢观看的电视频道或者节目类

型，如点播频道，回看频道或者直播频道等，节目类型如体育、电视剧、购物、少儿等等；

从付费方式维度分析用户是收费用户还是免费用户。

通过用户画像把握广电用户群体的特征和收视行为习惯模式，了解客户的实际特征和实际需求，并提供个性化、精准化和智能化的推荐服务。为用户提供一种更直接、更方便、更个性化的用户体验，以此来挽留客户、减少客户的流失。

1.2 项目目标

本项目根据广电用户基本信息及订单信息等数据，对这些数据进行预处理，删除无用字段信息。在处理好后的数据基础上构建用户画像，这些画像是用户的消费能力，及消费水平等用户挽留等级的重要体现。依据用户画像训练 svm 模型，进而得到可以预测用户是否需要挽留的机器学习模型。

1.3 项目数据

用户基本信息表，用户状态信息变更表，账单信息表，订单信息表用户收视行为信息表。如图 1，数据具体字段说明请见数据字段说明文档。






 media_index.csv	2018/11/16 13:43	Microsoft Excel ...	800,026 KB
 mediamatch_userevent.csv	2019/5/14 9:42	Microsoft Excel ...	242 KB
 mediamatch_usermsg.csv	2018/11/16 18:03	Microsoft Excel ...	13,547 KB
 mmconsume_billevents.csv	2018/11/16 13:41	Microsoft Excel ...	31,361 KB
 order_index.csv	2018/11/16 13:42	Microsoft Excel ...	139,186 KB

图 1 数据

1.4 项目周期

2 周

1.5 项目难度

★★★

2 项目任务

任务1. 明确项目需求与目标

任务2. 明确数据各表字段具体信息

任务3. 数据存储（mysql, hadoop, hive, hbase）

任务4. 数据探索

任务5. 数据预处理

任务6. 用户画像构建

任务7. Svm 算法模型构建（可扩展其他算法模型）

任务8. 学生项目成果提交

3 项目流程

- (1) 明确项目目标：项目经理对项目进行介绍展示，明确项目交付内容。
- (2) 学习前置知识：提供项目所需知识的云课堂课程，快速掌握项目前置知识。
- (3) 项目实践：在企业工程师带领下获取项目需求，动手做项目。
- (4) 项目验收：对项目成果进行验收，指导改进项目成果。
- (5) 信息入库：参训学生信息录入泰迪人才库，为优秀结业生引荐合作公司。

4 项目核心

- 数据存储及读取
- Spark 技术
- 根据数据及项目要求对数据进行探索
- 数据预处理
- 特征构建
- 模型训练，优化
- 模型评价

5 实现工具

Navicat/Hbase/Hive; Spark

6 实训对象

数学、统计学、计算机等相关专业学生。

7 前置知识

项目所需的所有前置知识，皆提供在线课程供学生免费学习。

第 1 模块：Spark 大数据技术基础

第 2 模块：大数据 Hive 数据仓库

第 3 模块：Hadoop 大数据开放基础

8 实训对应的就业岗位

8.1 就业岗位

数据分析师。

8.2 岗位分析

序号	岗位	主要业务工作	所需技能	相应课程设置
1	大数据工程师	数据处理、分析建模、撰写分析报告	Spark; Hadoop	Spark 大数据技术基础 Hadoop 大数据开放基础

9 实训收益

- 足不出校门即可获得实战技能。
- 全面实践数据分析流程，包括数据处理、数据探索、数据建模等。
- 掌握一定的挖掘技能和工具，体验一个实际项目的全过程。
- 获得企业实习证明，无需奔波解决实习需求。
- 获得 CBDA 大数据分析工程师证（初级），国际高水平认证。
- 参训学生信息录入泰迪人才库，为优秀结业生引荐合作公司。

10 证书

大数据分析工程师 基础级 技能证书

成

证书编号：CBDA-BDAE-11111111111111111111

2018年04月13日，通过认证考核，特发此证！

Big Data Analysis Engineer Level-1

heng

Certificate No. : CBDA-BDAE-11111111111111111111

Apr 13, 2018, through the certification examination



泰迪智能研究院国际培训中心
Tipdm Intelligent Institute International Training Center





11 附件一 前置课程课表

Spark 大数据技术基础	Hadoop 大数据开放基础
第 1 模块：Spark 简介及架构原理 课时 1：1.1.1 Spark 简介及发展历史介绍 06:36 课时 2：1.1.2 Spark 的特点 06:22	<ul style="list-style-type: none"> 课时 1：微班级开班仪式 05:42 第 1 模块：Hadoop 简介、核心及生态系统 第 1 节：Hadoop 简介

<p>课时 3 : 1.1.3 Spark 生态圈 11:22</p> <p>课时 4 : 1.1.4 Spark 的应用场景 03:18</p> <p>课时 5 : 1.2.1 搭建单机版 Spark 环境并运行 Pi 实例 04:57</p> <p>课时 6 : 1.2.2 搭建单机伪分布式 环境 12:00</p> <p>课时 7 : 1.2.3 搭建完全分布式环 境及启动、关闭、监控集群 23:43</p> <p>课时 8 : 1.3.1 Spark 架构 03:04</p> <p>课时 9 : 1.3.2 Spark 作业运行流程 05:45</p> <p>课时 10 : 1.3.3 Spark 核心数据集 RDD06:14</p> <p>课时 11 : 1.3.4 Spark 核心原理 08:53</p> <p>第 2 模块: Spark 编程基础: 常用 Transformations 和 Actions</p> <p>课时 12 : 2.1.1 从内存中已有数据 创建 RDD-Parallelize() 方法 06:15</p> <p>课时 13 : 2.1.2 从内存中已有数据 创建 RDD-makeRDD() 方法 05:48</p> <p>课时 14 : 2.1.3 从外部存储创建 RDD07:13</p> <p>课时 15 : 2.1.4 创建学生成绩数据 RDD03:52</p> <p>课时 16 : 2.2.1 使用 map() 转换数 据 04:59</p> <p>课时 17 : 2.2.2 使用 flatMap 转换 数据 04:24</p>	<ul style="list-style-type: none"> • 课时 2 : 1.1.1 Hadoop 简介 与发展历史 11:23 • 课时 3 : 1.1.2 Hadoop 的特 点 01:34 • 第 2 节: Hadoop 核心组件 • 课时 4 : 1.2.0 分布式公共 设施—Common00:37 • 课时 5 : 1.2.1 分布式文件 系统—HDFS09:29 • 课时 6 : 1.2.2 分布式计算 框架—MapReduce10:36 • 课时 7 : 1.2.3 集群资源管 理器—YARN01:43 • 第 3 节: Hadoop 生态系统 • 课时 8 : 1.3 Hadoop 生态系 统 02:54 • 第 4 节: Hadoop 应用场景 • 课时 9 : 1.4 Hadoop 应用场 景 01:16 • 第 2 模块: Hadoop 集群搭建 • 第 1 节: 安装配置虚拟机 • 课时 10 : 2.1.1 创建 Linux 虚拟机 17:40 • 课时 11 : 2.1.2 设置固定 IP07:26 • 课时 12 : 2.1.3 远程连接虚 拟机 07:25
--	--

<p>课时 18 : 2.2.3 使用 mapPartitions() 对分区操作 02:29</p> <p>课时 19 : 2.2.4 使用 sortBy() 排序 03:49</p> <p>课时 20 : 2.2.5 使用 Collect() 查询 01:21</p> <p>课时 21 : 2.2.6 使用 saveAsTextFile() 将 RDD 保存到 HDFS 中 03:13</p> <p>课时 22 : 2.2.7 使用 take() 方法查询某几个值 00:35</p> <p>课时 23 : 2.2.8 使用 count() 方法计算 RDD 中元素个数 00:39</p> <p>课时 24 : 2.2.9 查询学生成绩 Top507:55</p> <p>课时 25 : 2.3.1 使用 filter() 过滤 RDD02:23</p> <p>课时 26 : 2.3.2 使用 distinct() 进行去重 00:58</p> <p>课时 27 : 2.3.3 使用 union() 合并多个 RDD02:03</p> <p>课时 28 : 2.3.4 使用 subtract() 获取 RDD 之间的差集 02:56</p> <p>课时 29 : 2.3.5 使用 intersection() 找出 RDD 的交集 02:22</p> <p>课时 30 : 2.3.6 使用 cartesian() 进行集合笛卡尔积 02:58</p> <p>课时 31 : 2.3.7 查找单科成绩为 100 分的学生 03:46</p> <p>课时 32 : 2.4.1 键值对 RDD 简介 01:18</p>	<ul style="list-style-type: none"> • 课时 13 : 2.1.4 虚拟机在线安装软件 12:37 • 课时 14 : 【本地实训】 实训 1 创建及配置虚拟机 • 第 2 节: 安装 Java • 课时 15 : 2.2.1 在 windows 下安装 Java12:40 • 课时 16 : 2.2.2 在 Linux 下安装 Java04:09 • 课时 17 : 【在线实训】 实训 2 安装 Java • 课时 18 : 第一周作业 • 第 3 节: 搭建 Hadoop 完全分布式集群 • 课时 19 : 2.3.1 修改配置文件 24:01 • 课时 20 : 2.3.2 克隆虚拟机 15:31 • 课时 21 : 2.3.3 配置 SSH 免密码登录 10:20 • 课时 22 : 2.3.4 配置 ntp 时间同步服务&格式化 14:34 • 课时 23 : 2.3.5 启动关闭集群 11:15 • 课时 24 : 2.3.6 监控集群 21:31
---	---

<p>课时 33 : 2.4.2 创建键值对 RDD03:53</p> <p>课时 34 : 2.4.3 转换操作 Keys 与 Values01:54</p> <p>课时 35 : 2.4.4 转换操作 mapValues03:23</p> <p>课时 36 : 2.4.5 转换操作 groupBy()02:38</p> <p>课时 37 : 2.4.6 转换操作 groupByKey()03:24</p> <p>课时 38 : 2.4.7 转换操作 reduceByKey()06:48</p> <p>课时 39 : 2.4.8 汇总所有学生总成绩 05:08</p> <p>课时 40 : 2.5.1 使用 join()连接两个 RDD05:10</p> <p>课时 41 : 2.5.2 使用 zip 组合两个 RDD03:22</p> <p>课时 42 : 2.5.3 使用 combineByKey 合并相同键的值 15:02</p> <p>课时 43 : 2.5.4 使用 lookup 查找 指定键的值 00:39</p> <p>课时 44 : 2.5.5 计算每个学生平均 成绩 04:28</p> <p>课时 45 : 2.6.1 Json 文件的读取与 存储 15:42</p> <p>课时 46 : 2.6.2 csv 文件的读取与 存储 19:35</p> <p>课时 47 : 2.6.3 SequenceFile 与文 本文件读取和存储 07:33</p> <p>课时 48 : 2.6.4 文本文件的读取与 存储 03:32</p>	<ul style="list-style-type: none"> • 课时 25 : 【本地实训】实训 3 搭建 Hadoop 完全分布式集 群 • 课时 26 : 【本地实训】实训 4 编写 shell 脚本同步集群时 间 • 课时 27 : 【本地实训】实训 5 编写脚本来控制集群的启动 与关闭 • 第 3 模块: Hadoop 基础操作 • 第 1 节: 查看 Hadoop 集群的 基本信息 • 课时 28 : 3.1.1 查询集群的 存储系统信息 14:27 • 课时 29 : 3.1.2 查询集群的 计算资源信息 08:02 • 课时 30 : 【在线实训】实训 6 查看 Hadoop 集群的基本信 息 • 第 2 节: 上传文件到 HDFS • 课时 31 : 3.2.1 了解 HDFS 文件系统及其命令 10:05 • 课时 32 : 3.2.2 掌握 HDFS 的基本操作 22:44 • 课时 33 : 【在线实训】实训 7 上传文件到 HDFS 目录 • 课时 34 : 【在线实训】实训 8 查看 HDFS 上的文件内容
--	---

<p>课时 49 : 2.6.5 统计学生所有信息并存入文件 10:17</p> <p>第 3 模块: Spark 编程进阶: IDEA 配置、缓存、分区</p> <p>课时 50 : 3.1.1 下载安装 IntelliJ IDEA 安装 06:56</p> <p>课时 51 : 3.1.2 Scala 插件安装与使用 08:27</p> <p>课时 52 : 3.1.3 配置 Spark 运行环境及编写 Spark 程序基础知识 04:56</p> <p>课时 53 : 3.1.4 运行 Spark 单词计数程序实例 16:45</p> <p>课时 54 : 3.2.1 持久化(缓存) 10:38</p> <p>课时 55 : 3.2.2 数据分区 26:49</p> <p>课时 56 : 3.2.3 计算价格波动的幅度 24:53</p> <p>第 4 模块: 课程配套资料下载</p> <p>课时 57 : 配套资料</p>	<ul style="list-style-type: none"> • 第 3 节: 运行首个 MapReduce • 课时 35 : 3.3.1 了解 Hadoop 官方的示例程序包 08:58 • 课时 36 : 3.3.2 提交 MapReduce 任务集群运行 08:47 • 课时 37 : 【在线实训】实训 9 运行首个 MapReduce 任务 • 课时 38 : 【在线实训】实训 10 统计文件中所有单词的平均长度 • 第 4 节: 管理多个 MapReduce 任务 • 课时 39 : 3.4.1 查询多个 MapReduce 任务 04:41 • 课时 40 : 3.4.2 中断 MapReduce 任务 07:48 • 课时 41 : 3.4.3 如何用命令查看与中断任务运行情况及文件块信息 03:09 • 课时 42 : 【在线实训】实训 11 查询与中断 MapReduce 任务 • 课时 43 : 第二周作业 • 第 4 模块: MapReduce 编程入门 • 第 1 节: 使用 Eclipse 创建 MapReduce 工程
---	---

	<ul style="list-style-type: none"> • 课时 44 : 4.1.1 MapReduce 原理简介、任务介绍、Eclipse 安装及 MapReduce 环境配置原理 13:16 • 课时 45 : 4.1.2 配置 MapReduce 环境实操 04:38 • 课时 46 : 4.1.3 第一个 MapReduce 工程 05:55 • 课时 47 : 【在线实训】实训 12 使用 Eclipse 创建 MapReduce 工程 • 第 2 节: 通过源码初识 MapReduce 编程 • 课时 48 : 4.2.1 通俗理解 MapReduce 原理 10:25 • 课时 49 : 4.2.2 了解 MR 实现词频统计的执行流程 08:23 • 课时 50 : 4.2.3 读懂官方提供的 WordCount 源码-driver 运行流程 07:46 • 课时 51 : 4.2.4 内置键值对类型及 MapReduce 词频统计处理逻辑 12:12 • 课时 52 : 4.2.5 MapReduce 单词计数源码-打包运行 11:14 • 课时 53 : 4.2.6 Hadoop MapReduce-深入理解 08:17
--	--

	<ul style="list-style-type: none"> • 第 3 节：编程实现按日期统计访问次数 • 课时 54 : 4.3.1 访问次数统计其 MapReduce 实现思路与处理逻辑 09:27 • 课时 55 : 4.3.2 访问次数统计核心代码及实现 16:03 • 课时 56 : 【在线实训】实训 13 编程实现按日期统计访问次数 • 第 4 节：编程实现按访问次数排序 • 课时 57 : 4.4.1 访问次序排序分析思路与处理逻辑 06:23 • 课时 58 : 4.4.2 访问次序排序核心代码及其实现 09:36 • 课时 59 : 【在线实训】实训 14 编程实现按访问次数排序 • 课时 60 : 【在线实训】实训 15 获取成绩表的最高分记录 • 课时 61 : 【在线实训】实训 16 实现对两个文件中数据的合并与去重 • 课时 62 : 【在线实训】实训 17 统计 Hadoop 出现的次数 • 课时 63 : 第三周作业
--	--

	<ul style="list-style-type: none"> • 第 5 模块：MapReduce 编程进阶 • 第 1 节：筛选日志文件生成序列化文件 • 课时 64 : 5.1.1 MapReduce 输入格式 15:07 • 课时 65 : 5.1.2 MapReduce 输出格式 02:13 • 课时 66 : 5.1.3 日志文件筛选任务实现 25:15 • 课时 67 : 【在线实训】实训 18 筛选日志文件生成序列化文件 • 第 2 节：Hadoop Java API 读取序列化日志文件 • 课时 68 : 5.2.1 FileSystem API 管理文件夹 11:40 • 课时 69 : 5.2.2 FileSystem API 文件夹创建、删除实践 12:34 • 课时 70 : 5.2.3 FileSystem API 上传和下载 12:17 • 课时 71 : 5.2.4 FileSystem API 读写数据 15:03 • 课时 72 : 【在线实训】实训 19 Hadoop Java API 读取序列化日志文件
--	---

	<ul style="list-style-type: none"> • 第 3 节：优化日志文件统计程序 • 课时 73 : 5.3.1 自定义键值类型 22:54 • 课时 74 : 5.3.2 初步探索 Combiner13:07 • 课时 75 : 5.3.3 浅析 Partitioner11:42 • 课时 76 : 5.3.4 内置及自定义计数器 12:59 • 课时 77 : 【在线实训】实训 20 优化日志文件统计程序 • 第 4 节：Eclipse 提交日志文件统计程序 • 课时 78 : 5.4.1 传递参数 19:17 • 课时 79 : 5.4.2 Eclipse 自动打包并提交任务 10:44 • 课时 80 : 5.4.3 Hadoop 辅助类 ToolRunner 及统计任务实现 19:56 • 课时 81 : 【在线实训】实训 21 Eclipse 提交日志文件统计程序 • 课时 82 : 实训 22 统计全球每年的最高气温和最低气温 • 课时 83 : 实训 23 筛选气温在 15 到 25 度之间的数据
--	---

	<ul style="list-style-type: none">• 课时 84 : 实训 24 计算学生平均成绩• 课时 85 : 实训 25 QQ 好友推荐• 课时 86 : 第四周作业• 第 6 模块: 课程配套资料下载• 课时 87 : 课程配套资料
--	--

12 作业提交时间表

任务安排	内容	提交场所	提交时间
第一周	任务 1. 明确项目需求与目标 (5%)	泰迪云课堂	Day1 22:00
	任务 2.明确数据各表字段具体信息 (10%)		
	任务 3.数据存储 (15%)		Day2 22:00
	任务 4.数据探索 (20%)		
	任务 5.数据预处理 (40%)		Day5 22:00
第二周	任务 6.用户画像构建 (80%, 每个小任务为 5%进度)		Day1 22:00
	任务 6.1 消费内容		
	任务 6.2 电视消费水平		
	任务 6.3 宽带消费水平		
	任务 6.4 销售品名称		
	任务 6.5 宽带产品带宽		
	任务 6.6 业务水平		
	任务 6.7 电视入网程度		
	任务 6.8 宽带入网程度		
	任务 7.svm 模型构建 (90%)		Day3 22:00

	任务 8.项目成果提交（100%）	Day5 22:00
--	-------------------	------------

13 评分标准

注意：项目总分=总计*项目难度系数						
项目结果有 3 次修改机会，以最终提交成果为依据						
序号	评审项目	子项目	指 标	满分	分数	备注
1	项目进度	项目进度	在项目指定截止时刻的项目进度（如 90%，70%）	20		分数=20*项目进度
2	报告文档	内容完整性	与模板进行对比，是否各个步骤都在文档中体现出来	10		
		排版	与模板进行对比，是否按标准格式进行排版	10		
		内容质量	方案合理性 思路是否清晰 模型结果是否优异 创新性	30		
3	汇报PPT	内容完整性	与模板进行对比，是否各个步骤都在文档中体现出来	5		
		排版	与模板进行对比，是否按标准格式进行排版	5		
		内容质量	方案合理性 思路是否清晰 模型结果是否优异 创新性	5		
4	代码&中间数据	内容完整性	是否包含项目解决方案的各个环节对应的代码；是否有必要的中间数据。	10		

		代码规范 与设计	代码编写是否简洁、规范、高效；脚本分割与命名是否与模板对应。	5		
总计						
总分						

工作室邀请函

为了适应大数据与人工智能及发展的需求，顺应教育部提倡的深化校企合作的号召，更好的服务于广大数据分析爱好者，广东泰迪智能科技股份有限公司诚邀各高校相关专业老师、相关协会学会、俱乐部等组织合作成立“泰迪·智能工作室”，工作室以独立的模式运营，并以学生为中心成立，受泰迪科技监督且由其免费提供各种工作室所需资源的创新型数据智能工作室。



工作室旨在通过教育与产业之间的联动，实行“引进来，走出去”模式，引导学生学习数据科学与人工智能方法为导向，通过与企业的联系、合作、实践，激发学生的数据分析思维，全面推进数据分析与人工智能发展，提高大学生的数据分析素质，激发学生的创新创业精神，以实现创新型数据智能创业人才为培养目标。