# 北航数据挖掘课程Project 1

流媒体客户流失分析预测

## 背景介绍

说到流媒体，相信大家对中国的爱奇艺、优酷、腾讯视频，以及美国的亚马逊、Netflix、迪士尼等平台都不陌生。后疫情时代，视频流媒体的业务动态正在迅速变化。昔日以订阅用户增长为首要目标的时代已成过往。如今，所有流媒体业务都面临着严峻的市场现实——展示实际盈利能力的压力越来越大。内容制作和授权费用激增，订阅疲劳以及选择的奢侈让消费者一旦感觉不到价值或未被重视，便迅速取消服务。

客户流失（Churn）已成为流媒体服务提供商面临的一个关键问题。 2022年7月20日，网飞发布了第二季度财务报告，其“主战场”北美和加拿大地区的订阅数量从7458万减少到7328万，流失达130万。全球范围内的用户总计减少97万。[1]客户流失不仅会导致企业收入的减少，还会影响企业的市场竞争力和品牌声誉。图 1展示了2022 六月至 2023 七月，美国主要流媒体服务的客户流失率。

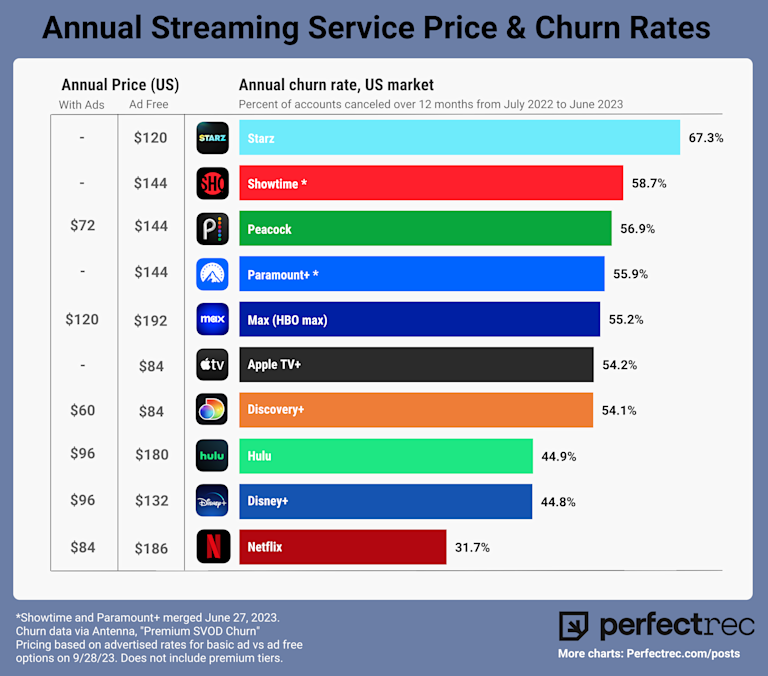


图 1 2022 六月至 2023 七月，美国主要流媒体服务的客户流失率。来源为Antenna。[2]

客户流失的定义通常指客户不再重复购买或终止服务，转而选择其他公司。客户流失的原因多种多样，包括产品或服务质量问题、价格因素、竞争对手的吸引力以及客户个人情况的变化等。为了降低客户流失率，企业需要深入了解客户的行为和需求，制定有效的客户保留策略。

此外，随着消费者对个性化体验的需求不断增加，个性化营销已成为企业提升竞争力的重要手段。个性化营销通过利用消费者数据和算法分析，提供满足其独特需求和偏好的定制化产品和服务。研究表明，个性化营销显著提高了客户满意度和忠诚度，尤其是在流媒体服务行业中，个性化推荐和定制化服务能够有效提升客户体验，从而降低流失率。

在这一背景下，我们选取了某流媒体流失数据集，通过数据挖掘技术，深入分析流媒体服务客户的流失模式、满意度水平和消费行为。通过对客户数据的全面分析，我们希望揭示客户流失的潜在因素，优化客户满意度，并制定个性化营销策略，以帮助企业提高客户保留率和市场竞争力。这不仅有助于流媒体服务提供商在激烈的市场竞争中脱颖而出，还能为行业的可持续发展提供数据驱动的决策支持。

## 项目目标

1. **数据预处理与统计分析**
   1. 数据预处理：数据集中包含的变量（如年龄、满意度评分等）存在缺失值，这为数据预处理提供了实际应用场景。通过填补缺失值、处理异常值和数据转换等方法，可以有效地清理和准备数据。
   2. 统计分析：数据集中的变量（如订阅时长、支持工单数量、满意度评分、月消费金额等）可以用于探索数据特征与客户流失之间的关系。通过统计分析方法（如相关性分析、分布分析等），可以识别出与客户流失相关的特征。
   3. 特征工程：基于统计分析结果，可以选择和构建对客户流失预测最有价值的特征。例如，可以创建新的特征（如客户活跃度、折扣利用率等），观察其是否可以提高模型的预测能力。
2. **客户流失预测模型开发**
   1. 数据挖掘算法：数据集中的变量（如年龄、性别、订阅时长、满意度评分、支持工单数量、月消费金额等）可以作为模型的输入特征。通过选择2个或2个以上的合适的数据挖掘算法，可以建立客户流失预测模型。
3. **模型性能评估**
   1. 评估指标：数据集中的“Churned”字段提供了客户是否流失的标签，可以用于评估模型的性能。常用的评估指标（如准确率、召回率、F1分数、ROC曲线下面积（AUC）等）可以全面评估模型的预测能力。
   2. 综合评估：通过多指标的综合评估，可以确保模型在预测客户流失方面的准确性和可靠性。例如，可以使用交叉验证方法来评估模型的稳定性和泛化能力。
4. **客户画像与个性化营销策略**
   1. 客户画像：数据集中的变量（如年龄、性别、订阅时长、满意度评分、支持工单数量、月消费金额等）可以用于构建客户画像。通过聚类分析或分类分析，可以将客户分为不同的群体，每个群体具有相似的特征和行为模式，分析并给出合理的个性化营销策略（自圆其说、合理即可）。

## 数据格式

本项目的数据集包含训练集 train.csv 和测试集 test.csv 两个文件（共5000条数据）[3]，存放在 dataset 文件夹下。模型应在训练集上进行训练，并在测试集上进行指标验证。

数据集的每行是一条独立的流媒体客户记录，共包含 12 个字段，各字段含义如下：

表 1 数据集字段说明

|  |  |
| --- | --- |
| 变量名 | 说明 |
| Customer\_ID | 客户的唯一标识符，用于区分不同的客户 |
| Age | 客户的年龄，部分记录可能存在缺失值 |
| Gender | 客户的性别，可能的取值包括但不限于 "Male"（男性）和 "Female"（女性） |
| Subscription\_Length | 客户的订阅时长，以月为单位 |
| Region | 客户所在的地区，可能的取值包括但不限于 "North"（北部）、"South"（南部）、"East"（东部）和 "West"（西部）。 |
| Payment\_Method | 客户使用的支付方式，可能的取值包括但不限于 "Credit Card"（信用卡）、"Debit Card"（借记卡）、"PayPal"（贝宝）和 "Bank Transfer"（银行转账）。 |
| Support\_Tickets\_Raised | 客户提交的支持工单数量 |
| Satisfaction\_Score | 客户满意度评分，范围通常在 0 到 10 之间，部分记录可能存在缺失值。 |
| Discount\_Offered | 提供给客户的折扣金额 |
| Last\_Activity | 客户最近一次与服务互动的天数。 |
| Monthly\_Spend | 客户每月的消费金额 |
| Churned | 客户是否流失的指示器，1 表示流失，0 表示未流失。 |

1. 数据集在本课程中期以北航网盘的形式发布。

## 注意事项

1. 本项目由助教提供baseline代码，与数据集一同在北航网盘发布，<https://bhpan.buaa.edu.cn/link/AAC1CC180EC9C8453B9BCCEFAED560D557>
2. 可使用python main.py命令在main.py所在文件夹下运行。
3. baseline的实现不是最优的。各小组实现结果应不低于或部分高于baseline。
4. 【提交格式】各小组结课时需提供项目的实现代码、实验报告、答辩PPT进行打包上传（zip格式），打包提交格式如图 2（见下页），提交链接<https://bhpan.buaa.edu.cn/link/AACEC41FE1795344C98248B155C2136702>
5. 实验报告按照《北航数据挖掘课程实验报告模板-2025春》模板撰写，链接<https://bhpan.buaa.edu.cn/link/AA434EA470CD5F44DB8A21B68AEB01DE7D>
6. 请注意两个大作业分开撰写和提交。

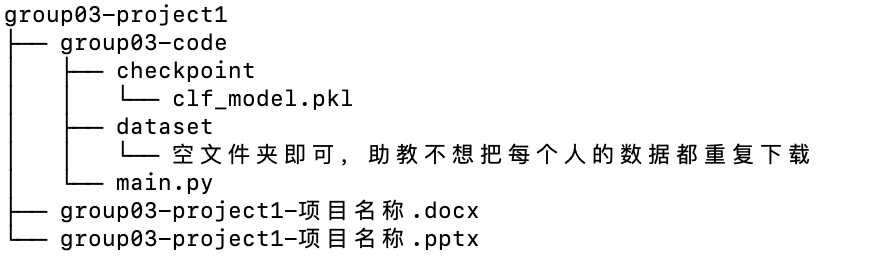


图 2 提交格式示例

1. 【代码规范】建议使用python语言完成项目。项目代码应具备结果可复现性，结果不能复现的、与报告汇报结果差距较大且不能合理说明理由的，按抄袭处理。
2. 【checkpoint】如果模型训练耗时较长，可以修改代码逻辑为"如果模型checkpoint已经存在，则直接加载模型，否则重新训练"，方便助教评测。
3. 【数据预处理】为避免在数据预处理过程中误操作导致的数据泄露，建议代码包含数据预处理算法，方便助教检查或有可能复现从原始数据到模型训练的全流程。
4. 项目代码和报告严禁抄袭，发现抄袭且不能合理说明理由的，一律按0分处理。
5. 项目报告严禁使用ChatGPT等AIGC写作工具直接生成[4]，发现明显AI写作痕迹的，酌情扣分。
6. 本文档将随课程进度更新，以最新版为准。更新通知会发到课程群。

编辑：陈一文

更新日期：2025年3月14日

## 参考文献

1. <https://www.thepaper.cn/newsDetail_forward_19093197>
2. <https://www.perfectrec.com/posts/chart-streaming-service-churn-rates>
3. 本项目涉及数数据处理修改自<https://www.kaggle.com/datasets/akashanandt/streaming-service-data>
4. 郑雯雯等. 学术出版中AIGC使用边界指南. 中国科学技术信息研究所, 2023.