# **基于PySpark的客户流失研究**案例教学指导手册

**教学适用的课程：**《大数据架构与设计》

**教学适用的专业：**电子信息

**教学目的与用途：**理解商业背景下的大数据应用；掌握机器学习、数据分析相关理论；掌握大数据处理与分析的基本流程。

**教学内容：**

1. 理论依据：本案例为基于PySpark的客户流失研究。客户流失分析可以更好地了解客户流失的根本原因，以减少客户流失和提高产品销售额。通过分析，企业可以找出是什么原因导致客户离开，谁有可能离开，以及自己能做些什么来留住客户。案例使用了常用的数据挖掘与机器学习方法，进行数据处理与分类预测，从而实现客户流失分析。
2. 涉及知识点：客户流失分析、二分类预测、逻辑回归、梯度提升树、超参调优、交叉验证、可视化、PySpark
3. 分析路径：首先介绍数据集相关字段含义。然后通过数据预处理、特征工程、模型训练实现了逻辑回归和梯度提升树方法的用户流失二分类预测，并对模型使用AUC指标和PR指标进行评估。由于第一个模型的参数往往不是最佳的，因此又使用K-Fold交叉验证进行超参调优，从而进一步提升模型评估的AUC指标和PR指标。最后可视化展示了特征与流失用户关系图以及特征重要图，进一步凸显客户流失领域的重要特征。

**启发思考题：**

超参调优除了K-Fold交叉验证你还知道哪些其他方法？（参考答案：比如网格搜索）

**建议课堂计划：**

1. 时间安排：3课时
2. 学生学习准备：了解常见的数据预处理、分类预测方法、PySpark的使用。
3. 分组及讨论内容：分组讨论启发思考题。
4. 案例开场白：客户流失率是指客户的流失数量与全部消费产品或服务客户的数量的比例，客户流失率是判断客户流失的主要指标。通过客户流失分析可以更好地了解客户流失的根本原因，以减少客户流失和提高产品销售额。因此客户流失对于企业有着至关重要的作用，这节课我们将一起探讨大数据如何分析客户流失问题。
5. 结束总结：这节课我们以客户流失分析为例，使用PySpark进行数据预处理、特征工程、模型训练，从而实现了用户流失的二分类预测，并使用AUC指标和PR指标进行评估。我们还使用K-Fold交叉验证进行超参调优，进一步提升模型评估指标。最后通过特征与流失用户关系图以及特征重要图进行进一步可视化分析。相信通过这一节课，大家对PySpark的使用有了一定的了解，课后感兴趣的同学也可以继续深入了解，尝试使用更多的方法解决本堂课探讨的问题。
6. 案例引导建议：在教师简要分析应用场景后，鼓励学生自主考虑如何使用所学知识解决该场景问题。

**参考文献：**

[1] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.

参考内容：逻辑回顾、梯度提升树、超参调优、交叉验证法、AUC指标、PR指标

**其他教学支持材料：**

1. 本案例提供配套的PPT、视频、数据集与代码等，发布于Github，链接为：https://github.com/Wanghui-Huang/CQU\_bigdata。

2. 本案例涉及到数据处理以及机器学习算法，使用PySpark进行实现。