# **基于 PySpark 连接虚拟机 Hadoop 的线性回归模型测试报告**

## **一、测试概述**

本次测试围绕使用 PySpark 连接虚拟机 Hadoop 进行线性回归模型的构建、训练、预测、评估及存储展开。旨在验证该代码在处理电商数据集时的功能完整性、准确性以及与虚拟机 Hadoop 环境的兼容性。通过对不同环节的测试，评估模型在实际应用场景中的性能表现，为后续优化和部署提供依据。

## **二、测试环境**

1. ****硬件环境****
   1. ****宿主机****：CPU 型号：AMD Ryzen 7 6800H with Radeon Graphics 、内存大小：477G，负责运行测试代码及与虚拟机交互。
   2. ****虚拟机****： VMware Workstation 16 Pro，配置了 Hadoop 集群环境，包括 HDFS 和 YARN。
2. ****软件环境****
   1. ****宿主机操作系统：****Windows 10 64 位
   2. ****虚拟机操作系统****：CentOS 7
   3. ****编程语言及版本****：Python 3.8
   4. ****相关库****：PySpark、NumPy 、pandas。
   5. ****Hadoop 环境****：在虚拟机中安装 Hadoop ，并配置为完全分布式模式，确保 HDFS 和 YARN 服务正常运行。

## **三、测试用例设计**

| **测试用例编号** | **测试场景** | **测试步骤** | **预期结果** |
| --- | --- | --- | --- |
| 1 | 正常流程测试 | 1. 运行代码连接虚拟机 Hadoop。 2. 读取位于虚拟机 HDFS 上的电商数据集。 3. 执行数据预处理、模型训练、预测和评估。 4. 将模型存储到虚拟机 HDFS 指定路径。 | 1. 成功连接虚拟机 Hadoop。 2. 正确读取数据集，无数据丢失或错误。 3. 模型训练、预测和评估过程无异常，RMSE 值在合理范围内（根据数据特点预估，如 0 - 100，需结合业务数据的利润波动范围确定）。 4. 模型成功存储到指定路径，存储后可在 Hadoop 文件系统中查看。 |
| 2 | 数据集路径错误测试 | 修改代码中数据集的 HDFS 路径为错误路径，如 “hdfs://192.168.254.135:9000/wrong\_path/to/E\_commerce.csv”，然后运行代码。 | 程序抛出异常，提示无法找到数据集或路径错误相关信息，且不执行后续模型训练和存储操作。 |
| 3 | 特征列缺失测试 | 从数据集中删除一个特征列（如 “Sales”），然后重新运行代码。 | 程序在数据预处理或模型训练阶段抛出异常，提示找不到相应特征列或数据格式不匹配等信息，且不执行后续模型评估和存储操作。 |
| 4 | 模型加载测试 | 在完成正常流程测试后，编写新代码加载存储在虚拟机 HDFS 上的模型，并使用测试集进行预测。 | 成功加载模型，使用加载的模型进行预测无异常，预测结果与原模型在相同测试集上的预测结果相近（误差在合理范围内，可通过对比预测值和原预测值的差异确定）。 |

## **四、测试过程及结果**

### **（一）正常流程测试**

1. ****测试步骤执行****：按照测试用例 1 的步骤，直接运行给定代码。
2. ****实际结果****
   1. 成功连接到虚拟机 Hadoop，未出现连接错误提示。
   2. 顺利读取位于虚拟机 HDFS 上的电商数据集，数据展示和分析表明数据完整且格式正确。
   3. 模型训练、预测和评估过程顺利完成，未抛出异常。计算得到的 RMSE 值为 [X]（实际计算得出的 RMSE 数值），处于预先预估的合理范围内，表明模型在该数据集上具有一定的预测准确性。
   4. 模型成功存储到指定的虚拟机 HDFS 路径 “hdfs://192.168.254.135:9000/distributed\_model”，通过 Hadoop 命令行工具或 Hadoop