银行贷款违约预测

背景

在借贷交易中，银行和其他金融机构通常提供资金给借款人，期望借款人能够按时还款本金和利息。然而，由于各种原因，有时借款人可能无法按照合同规定的方式履行还款义务，从而导致贷款违约。本次实验以银行贷款违约为背景，选取了约30万条贷款信息 ， 包含在application_data.csv文件中，数据描述包含在columns_description.csv文件夹中。

数据来源：https://www.kaggle.com/datasets/mishra5001/credit-card/data

任务

任务一

1、编写 Spark 程序，统计application_data.csv中所有用户的贷款金额**AMT_CREDIT** 的分布情况。

以 10000 元为区间进行输出。输出格式示例：

```
((20000,30000),1234)
```

表示20000到30000元之间（包括20000元，但不包括30000元）有1234条记录。

2、编写Spark程序，统计application_data.csv中客户贷款金额**AMT_CREDIT** 比客户收入**AMT_INCOME_TOTAL**差值最高和最低的各十条记录。

输出格式：

```
<SK_ID_CURR><NAME_CONTRACT_TYPE><AMT_CREDIT><AMT_INCOME_TOTAL>, <差值>
```

差值=AMT_CREDIT-AMT_INCOME_TOTAL

任务二

基于Hive或者Spark SQL对application_data.csv进行如下统计：

1、统计所有男性客户（CODE_GENDER=M）的小孩个数（CNT_CHILDREN）类型占比情况。

输出格式为：

```
<CNT_CHILDREN>，<类型占比>
```

例：

```
0，0.1234
```

表示没有小孩的男性客户占总男性客户数量的占比为0.1234。

2、统计每个客户出生以来每天的平均收入（avg_income）=总收入（AMT_INCOME_TOTAL）/出生天数（DAYS_BIRTH)，统计每日收入大于1的客户，并按照从大到小排序，保存为csv。

输出格式：

```
<SK_ID_CURR>, <avg_income>
```

任务三

根据给定的数据集，基于Spark MLlib 或者Spark ML编写程序对贷款是否违约进行分类，并评估实验结果的准确率。可以训练多个模型，比较模型的表现。

说明：

1、该任务可视为一个“二分类”任务，因为数据集只存在两种情况，违约（Class=1）和其他（Class=0）。

2、可根据时间特征的先后顺序按照8：2的比例将数据集application_data.csv拆分成训练集和测试集，时间小的为训练集，其余为测试集；也可以按照8：2的比例随机拆分数据集。最后评估模型的性能，评估指标可以为accuracy、f1-score等。

3、基于数据集application_data.csv，可以自由选择特征属性的组合，自行选用一种或多种分类算法对目标属性**TARGET**进行预测。

提交方式

提交git仓库地址或者相关文件的zip包。实验报告应包括设计思路、运行结果和可能的改进之处等。