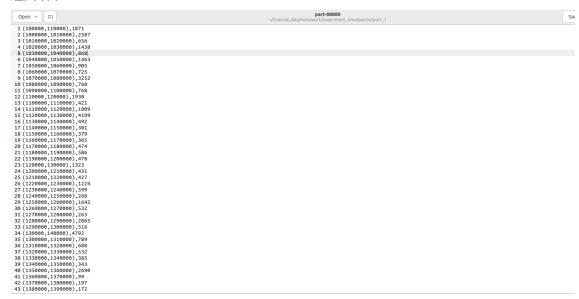
银行贷款违约预测

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。

- 首先将所有的AMT_CREDIT都转化为整型
- 新建一列"CREDIT_RANGE",用于存放AMT_CREDIT所属区间,形式为((AMT_CREDIT // 10000) * interval,(AMT_CREDIT// 10000+1) *10000)
- 根据CREDIT_RANGE对数据进行分组,并统计每一组的个数
- 返回结果



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。

- 新建一列DIFFERENCE,用于存放AMT_CREDIT和AMT_INCOME_TOTAL的差值
- 根据DIFFERENCE进行降序排列,取前十个记录,这是差值最高的十条记录
- 根据DIFFERENCE进行升序排列,取前十个记录,这是差值最低的十条记录
- 将结果合并为一个数据框,并选择要保留的列进行输出

	Α	В	C	D	E	F	G	Н	1	J	
1	SK ID ÇURR	NAME_CONTRACT_TYPE	AMT_CREDIT	AMT_INCOME_TOTAL	DIFFERENCE						
2	433294	Cash loans	4050000	405000	3645000						
3	210956	Cash loans	4031032.5	430650	3600382.5						
4	434170	Cash loans	4050000	450000	3600000						
5	315893	Cash loans	4027680	458550	3569130						
6	238431	Cash loans	3860019	292050	3567969						
7	240007	Cash loans	4050000	587250	3462750						
8	117337	Cash loans	4050000	760846.5	3289153.5						
9	120926	Cash loans	4050000	783000	3267000						
10	117085	Cash loans	3956274	749331	3206943						
11	228135	Cash loans	4050000	864900	3185100						
12	114967	Cash loans	562491	117000000	-116437509						
13	336147	Cash loans	675000	18000090	-17325090						
14	385674	Cash loans	1400503.5	13500000	-12099496.5						
15	190160	Cash loans	1431531	9000000	-7568469						
16	252084	Cash loans	790830	6750000	-5959170						
17	337151	Cash loans	450000	4500000	-4050000						
18	317748	Cash loans	835380	4500000	-3664620						
19	310601	Cash loans	675000	3950059.5	-3275059.5						
20	432980	Cash loans	1755000	4500000	-2745000						
21	157471	Cash loans	953460	3600000	-2646540						
22											
22											

1. 基于Spark SQL,统计所有男性客户(CODE_GENDER=M)的小孩个数(CNT_CHILDREN)类型占比情况

- 先将原始数据中CODE_GENDER=M的部分筛选出来,记作gender_filtered
- 然后计算gender_filtered中CNT_CHILDREN的分布

```
children_count = gender_filtered.groupBy("CNT_CHILDREN").count()
```

- 根据gender_filtered可以得出所有男性客户的总人数
- 在children_count新建一列ratio,值为count/总人数,结果保留4位小数
- 根据CNT_CHILDREN进行升序排列,输出结果

```
1 0,0.6693
2 1,0.2157
3 2,0.0991
4 3,0.0138
5 4,0.0016
6 5,0.0003
7 6,0.0001
8 7,0.0
9 8,0.0
10 9,0.0
11 11,0.0
12 14,0.0
```

- 2. 基于Spark SQL ,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
 - 在原始数据框新增一列avg_income,表示每个客户出生以来每天的平均收入
 - 筛选出avg_income>1的部分,并根据avg_income降序排列
 - 选择需要保存的列,输出

1 SK ID CURR avg income 2 114967 9274.6730082343 3 336147 1146.21051961284 4 385674 996.236440115121 5 190160 547.945205479452 6 219563 417.517164594544 7 310601 373.634080590238 8 157471 36.43680590238 8 157471 36.43618022827 9 252043 348.999534667287 10 199821 269.654841802493 11 337151 243.757109582363 12 141198 243.623676612127 13 429258 241.659394508962 14 196091 240.761877585961 15 317748 240.448837830617 16 432980 239.565587734242 17 217276 235.320484087853 18 445335 234.308435103664 19 387126 230.465320456641 20 304300 223.583968201391 21 123587 207.249674902471 22 399467 195.921921486104 23 441639 192.455735180908 24 440768 192.04096873998 25 225210 188.75381833737 26 206341 186.00165334803 27 134526 183.68846364384 29 336135 177.464788732394 30 111903 174.1515188859955 31 1296498 172.005198379329 32 194130 172.005198379329 33 1 11903 174.1515188859955 31 1 296498 172.005198379329 32 194130 172.005198379329		Α	В	C	D	E	F	G	Н	1	J
3 336147 1146.21051961284 4 385674 996.236440115121 5 190160 547.945205479452 6 219563 417.517164594544 7 310601 373.634080590238 8 157471 360.432519022827 9 252084 348.999534667287 10 199821 269.65641802493 11 337151 243.757109582363 12 141198 243.623676612127 13 429258 241.6589394508962 14 196091 240.761877585961 15 317748 240.448837830617 16 432980 239.565587734242 17 217276 235.320484087853 18 445335 234.308435103664 19 387126 230.465320456541 20 304300 223.583968201391 21 123587 207.249674902471 22 399467 195.921921486104 23 441639 192.455735180908 24 440768 192.04058733998 25 225210 188.75388133737 26 206341 186.00165334803 27 134526 183.6884638438 30 111903 174.185128859995 31 269498 177.6075193999 30 11903 174.185128859995 31 269498 177.464788732394 30 111903 174.185128859995 31 269498 177.464788732394 30 111903 174.185128859995 31 269498 172.602739726027 32 194130 172.005198379329	1	SK ID ÇURR	avg_income								
4 385674 996.236440115121 5 190160 547.945205479452 6 219563 417.517164594544 7 310601 373.634080590238 8 157471 360.432519022827 9 252084 348.99953467287 10 199821 269.654841802493 11 337151 243.757109582363 12 141198 243.623676612127 13 429258 241.659394508962 14 196091 240.761877585961 15 317748 240.448837830617 16 432980 239.565587734242 17 217276 235.320484087853 18 445335 234.308435103664 19 387126 230.465320456541 20 304300 223.583968201391 21 123587 207.249674902471 22 399467 195.921921486104 23 441639 192.455735180908 24 40768 192.040968739998 25 225210 188.75388133737 26 2063	2	114967	9274.67300832343								
5 190160 547.945205479452 6 219563 417.517164594544 7 310601 373.634080590238 8 157471 360.432519022827 9 252084 348.999534667287 10 199821 269.654841802493 11 337151 243.757109582363 12 141198 243.623676612127 13 429258 241.659394508962 14 196091 240.761877585961 15 317748 240.448837830617 16 432980 293.956589734242 17 217276 235.320484087853 18 445335 234.308435103664 19 387126 230.465320456541 20 304300 223.583968201391 21 123587 207.249674902471 22 399467 159.921921486104 23 441639 192.040968739998 24 440768 192.040968739998 25 225210 188.75388133737 26 206341 186.00165334803 27 134526 183.688464364438 28 214063 180.082716551444 29 336135 177.468788732394 30	3	336147	1146.21051961284								
6 219563 417.517164594544 7 310601 373.634080590238 8 157471 360.432519022827 9 252084 348.999534667287 10 199821 269.654841802493 11 337151 243.757109582363 12 141198 243.623676612127 13 429258 241.659394508962 14 196091 240.761877585961 15 317748 240.448837830617 16 432980 239.565587734242 17 217276 235.320484087853 18 445335 234.308435103664 19 387126 230.465320456541 20 304300 223.583968201391 21 123587 207.249674902471 22 399467 195.921921486104 23 441639 192.455735180908 24 440768 192.040968739998 25 225210 188.75388133737 26 206341 186.00165334803 27 134526 183.688464364438 28 214063 180.082716551444 29 336135 177.464788732394 30 111903 174.135128859995 31 269498 172.602739726027 32 194130 172.005198379329	4	385674	996.236440115121								
7 310601 373.634080590238 8 157471 360.432519022827 9 252084 348.999534667287 9 252084 348.999534667287 9 10 199821 269.654841802493 11 337151 243.757109582363 12 141198 243.623676612127 13 429258 241.659394508962 14 196091 240.761877585961 15 317748 240.448837830617 16 432980 239.565587734242 17 217276 235.320484087853 18 445335 234.308435103664 19 387126 230.465320456541 20 304300 223.583968201391 21 123587 207.249674902471 22 399467 195.921921486104 23 441639 192.2455735180908 24 440768 192.040968739998 25 225210 188.75388133737 26 206341 186.00165334803 27 134526 183.688464364488 29 336135 177.464788732394 30 111903 174.135128859995 31 269498 172.002193879329 194130 172.005198379329 31 269498 172.002193879329 31 269498 172.002193793929	5	190160	547.945205479452								
8 157471 360.432519022827 9 252084 348.999534667287 10 199821 269.654841802493 11 337151 243.757199582363 12 141198 243.623676612127 13 429258 241.659394508962 14 196091 240.761877585961 15 317748 240.448837830617 16 432980 239.565587734242 17 217276 235.320484087853 18 445335 234.308435103664 19 387126 230.465320456541 20 304300 223.583968201391 21 123587 207.249674902471 22 399467 195.921921486104 23 441639 192.455735180908 24 440768 192.04098739998 25 225210 188.75388133737 26 206341 186.00165334803 27 134526 183.688464364438 28 214063 180.082716551444 29 336135 177.464788732394 30 111903 174.1351288559995 31 269498 172.602739726027 32 194130 172.005198379329	6	219563	417.517164594544								
9 252084 348.999534667287 10 199821 269.654841802493 11 337151 243.757109582363 12 141198 243.623676612127 13 429258 241.659394508962 14 196091 240.761877585961 15 317748 240.448837830617 16 432980 239.565587734242 17 217276 235.320484087853 18 445335 234.308435103664 19 387126 230.465320456541 20 304300 223.583968201391 21 123587 207.249674902471 22 399467 195.921921486104 23 441639 192.455735180908 24 440768 192.04096873998 25 225210 188.75388133737 26 206341 186.00165334803 27 134526 183.68846364438 28 214063 180.082716551444 29 336135 177.464788732394 30 111903 174.135128859995 31 1269498 172.602739726027 32 194130 172.005198379329	7	310601	373.634080590238								
10 199821 269.654841802493 11 337151 243.757109582363 12 141198 243.623676612127 13 429258 241.659394508962 14 196091 240.761877585961 15 317748 240.448837830617 16 432980 239.565587734242 17 217276 235.320484087853 18 445335 234.308435103664 19 387126 230.465320456541 20 304300 223.583968201391 21 123587 207.249674902471 22 399467 195.921921486104 23 441639 192.455735180908 24 440768 192.040968739998 25 225210 188.75388133737 26 206341 186.00165334803 27 134526 183.688464364438 28 214063 180.082716551444 29 336135 174.135128859995 31 269498 172.602739726027 32 194130 172.005198379329	8	157471	360.432519022827								
11 337151 243.757109582363 12 141198 243.623676612127 13 429258 241.659394508962 14 196091 240.761877585961 15 317748 240.448837830617 16 432980 239.565587734242 17 217276 235.3204840878853 18 445335 234.308435103664 19 387126 230.465320456541 20 304300 223.583968201391 21 123587 207.249674902471 22 399467 195.921921486104 23 441639 192.455735180908 24 440768 192.040968739998 25 225210 188.75388133737 26 206341 186.00165334803 27 134526 183.688464364438 28 214063 180.082716551444 29 336135 177.464788732394 30 111903 174.135128859995 31 269498 172.602739726027 32 194130 172.005198379329	9	252084	348.999534667287								
12 141198 243.623676612127 13 429258 241.659394509962 14 196091 240.761877585961 15 317748 240.448837830617 16 432980 239.565587734242 17 217276 235.320484087853 18 445335 234.308435103664 19 387126 230.455320456541 20 304300 223.583968201391 21 123587 207.249674902471 22 399467 195.921921486104 23 441639 192.455735180908 24 440768 192.040968739998 25 225210 188.75388133737 26 206341 186.00165334803 27 134526 183.688464364438 28 214063 180.082716551444 29 336135 177.464788732394 30 111903 174.135128859995 31 269498 172.602739726027 32 194130 172.005198379329	10	199821	269.654841802493								
13 429258 241.659394508962 14 196091 240.761877585961 15 317748 240.448837830617 16 432980 239.565587734242 17 217276 235.320484087853 18 445335 234.308435103664 19 387126 230.465320456541 20 304300 223.583968201391 21 123587 207.249674902471 22 399467 195.921921486104 23 441639 192.455735180908 24 440768 192.040968739998 25 225210 188.7538133737 26 206341 186.00165334803 27 134526 183.688464364438 28 214063 180.082716551444 29 336135 177.464788732394 30 111903 174.135128859995 31 269498 172.602739726027 32 194130 172.005198379329	11	337151	243.757109582363								
14 196091 240.761877585961 15 317748 240.448837830617 16 432980 239.565587734242 17 217276 235.320484087853 18 445335 234.308435103664 19 387126 230.465320456541 20 304300 223.583968201391 21 123587 207.249674902471 22 399467 195.921921486104 23 441639 192.455735180908 24 440768 192.040968739998 25 225210 188.75388133737 26 206341 186.00165334803 27 134526 183.688464364438 28 214063 180.082716551444 29 336135 177.464788732394 30 111903 174.135128859995 31 269498 172.602739726027 32 194130 172.005198379329	12	141198	243.623676612127								
15 317748 240.448837830617 16 432980 239.565587734242 17 217276 235.320484087853 18 445335 234.308435103664 19 387126 230.465320456541 20 304300 223.583968201391 21 123587 207.249674902471 22 339467 195.921921486104 23 441639 192.455735180908 24 440768 192.040968739998 25 225210 188.75388133737 26 206341 186.00165334803 27 134526 183.688464364438 28 214063 180.082716551444 29 336135 177.464788732394 30 111903 174.135128859995 31 1269498 172.602739726027 32 194130 172.005198379329	13	429258	241.659394508962								
16 432980 239.565587734242 17 217276 235.320484087853 18 445335 234.308435103664 19 387126 230.465320456541 20 304300 223.583968201391 21 123587 207.249674902471 22 399467 195.921921486104 23 441639 192.455735180908 24 440768 192.040968739998 25 225210 188.75388133737 26 206341 186.00165334803 27 134526 183.688464364438 28 214063 180.082716551444 29 336135 177.464788732394 30 111903 174.135128859995 31 269498 172.602739726027 32 194130 172.005198379329	14	196091	240.761877585961								
17 217276 235.320484087853 18 445335 234.308435103664 19 387126 230.465320456541 20 304300 223.583968201391 21 123587 207.249674902471 22 399467 195.921921486104 23 441639 192.0455735180908 24 440768 192.040968739998 25 225210 188.75388133737 26 206341 186.00165334803 27 134526 183.688464364438 28 214063 180.082716551444 29 336135 177.464788732394 30 111903 174.135128859995 31 269498 172.602739726027 32 194130 172.005198379329 30 194130 172.005198379329 30 194130 172.005198379329 30 194130 172.005198379329 30 30 194130 172.005198379329 30 30 194130 172.005198379329 30 </td <td>15</td> <td>317748</td> <td>240.448837830617</td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td>	15	317748	240.448837830617								
18 445335 234.308435103664 19 387126 230.465320456541 20 304300 223.583968201391 21 123587 207.249674902471 22 399467 195.921921486104 23 441639 192.455735180908 24 440768 192.040968739998 25 225210 188.75388133737 26 206341 186.00165334803 27 134526 183.688464364438 28 214063 180.082716551444 29 336135 177.464788732394 30 111903 174.135128859995 31 269498 172.602739726027 32 194130 172.005198379329	16	432980	239.565587734242								
19 387126 230.465320456541 20 304300 223.583968201391 21 123587 207.249674902471 22 399467 195.921921486104 23 441639 192.455735180908 24 440768 192.040968739998 25 225210 188.75388133737 26 206341 186.00165334803 27 134526 183.688464364438 28 214063 180.082716551444 29 336135 177.464788732394 30 111903 174.135128859995 31 269498 172.602739726027 32 194130 172.005198379329	17	217276	235.320484087853								
20 304300 223.583968201391 21 123587 207.249674902471 22 399467 195.921921486104 23 441639 192.455735180908 24 440768 192.040968739998 25 225210 188.75388133737 26 206341 186.00165334803 27 134526 183.688464364438 28 214063 180.082716551444 29 336135 177.464788732394 30 111903 174.135128859995 31 269498 172.602739726027 32 194130 172.005198379329	18	445335	234.308435103664								
21 123587 207.249674902471 22 399467 195.921921486104 23 441639 192.455735180908 24 440768 192.040968739998 25 225210 188.75388133737 26 206341 186.00165334803 27 134526 183.688464364438 28 214063 180.082716551444 29 336135 177.464788732394 30 111903 174.135128859995 31 269498 172.602739726027 32 194130 172.005198379329	19	387126	230.465320456541								
22 399467 195.921921486104 23 441639 192.455735180908 24 440768 192.040968739998 25 225210 188.75388133737 26 206341 186.00165334803 27 134526 183.688464364438 28 214063 180.082716551444 29 336135 177.464788732394 30 111903 174.135128859995 31 269498 172.602739726027 32 194130 172.005198379329	20	304300	223.583968201391								
23 441639 192.455735180908 24 440768 192.040968739998 25 225210 188.75388133737 26 206341 186.00165334803 27 134526 183.688464364438 28 214063 180.082716551444 29 336135 177.464788732394 30 111903 174.135128859995 31 269498 172.602739726027 32 194130 172.005198379329	21	123587	207.249674902471								
24 440768 192.040968739998 25 225210 188.75388133737 26 206341 186.00165334803 27 134526 183.688464364438 28 214063 180.082716551444 29 336135 177.464788732394 30 111903 174.135128859995 31 269498 172.602739726027 32 194130 172.005198379329	22	399467	195.921921486104								
25 225210 188.75388133737 26 206341 186.00165334803 27 134526 183.688464364438 28 214063 180.082716551444 29 336135 177.464788732394 30 111903 174.135128859995 31 269498 172.602739726027 32 194130 172.005198379329	23	441639	192.455735180908								
26 206341 186.00165334803 27 134526 183.688464364438 28 214063 180.082716551444 29 336135 177.464788732394 30 111903 174.135128859995 31 269498 172.602739726027 32 194130 172.005198379329	24	440768	192.040968739998								
27 134526 183.688464364438 28 214063 180.082716551444 29 336135 177.464788732394 30 111903 174.135128859995 31 269498 172.602739726027 32 194130 172.005198379329	25	225210									
28 214063 180.082716551444 29 336135 177.464788732394 30 111903 174.135128859995 31 269498 172.602739726027 32 194130 172.005198379329	26	206341	186.00165334803								
29 336135 177.464788732394 30 111903 174.135128859995 31 269498 172.602739726027 32 194130 172.005198379329	27	134526	183.688464364438								
30 111903 174.135128859995 31 269498 172.602739726027 32 194130 172.005198379329											
31 269498 172.602739726027 32 194130 172.005198379329											
32 194130 172.005198379329	30	111903									
22 401111 100 750010700100											
33 431111 100.759310728182	33	431111	166.759310728182								

任务三 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。

1. 数据集分割

- 首先将原始数据清洗,将带有缺失值的行全部删除
- 利用sklearn库中的train_test_split方法对清洗后的数据集进行随机分割,训练集与测试集的大小之比为4: 1

2. 特征工程

- 对文本和离散特征进行 StringIndexer 和 OneHotEncoder 处理
- 过滤掉具有大量唯一值的列,因为这些特征大概率不会给模型带来泛化能力上的提升
- 将所有处理过的特征组合成一个向量, 便于后续模型训练
- 特征标准化

ps:模型实列化之后,即可与上述步骤组合成Pipeline,方便模型训练与预测

3. 不同模型的选择与性能表现

• 直接从pyspark.ml.classification 中选择想要的模型

逻辑回归

Accuracy: 0.9194510836869746 F1 Score: 0.8809446106263878

随机森林

Accuracy: 0.9195323805342829 F1 Score: 0.8809851893362669

支持向量机

Accuracy: 0.9194510836869746 F1 Score: 0.8809446106263878

可以看到这三个分类模型在该数据集上的性能都是差不多的。