Spark + Python: DO Big Data Analytics & ML

Course Apply Project: Problem Statement

This project corresponds to a real life scenario found in Credit Card companies. Customers who own credit cards are expected to pay off their balances monthly. But, they do default (not pay), which forces the bank into financial situations. Banks want to know which customer would possibly default in the future, so they can take necessary actions (such as closing their card, reducing their spending limits etc.). This problem involves a specific bank who wants to analyze their customer's payment patterns and narrow down to cases where they are most likely to default.

This problem has a dataset that contains information about Credit Card customers for the past 6 and a set of questions that the bank has. Your assignment is to analyze the data and come up with answers to these questions using Apache Spark.

The dataset is in a file credit-card-default-1000.csv. The file contains the following columns

Column Name	**Description**
CUSTID	Unique Customer ID
LIMIT_BAL	Maximum Spending Limit for the customer
SEX	Sex of the customer. Some records have M and F to indicate sex. Some records have 1 (Male) and 2 (Female)
EDUCATION	Education Level of the customer. The values are 1 (Graduate), 2 (University), 3 (High School) and 4 (Others)
MARRIAGE	Marital Status of the customer. The values are 1 (Single), 2 (Married) and 3 (Others)
AGE	Age of the customer
PAY_1 to PAY_6	Payment status for the last 6 months, one column for each month. This indicates the number of months (delay) the customer took to pay that month's bill
BILL_AMT1 to BILL_AMT6	The Billed amount for credit card for each of the last 6 months.
PAY_AMT1 to PAY_AMT6	The actual amount the customer paid for each of the last 6 months
DEFAULTED	Whether the customer defaulted or not on the 7th month. The values are 0 (did not default) and 1 (defaulted)

Problems to Solve and Questions to Answer.

PR#01: Do Data cleansing and enhancements as required to solve the problem. The dataset does have problems that you need to find out and fix. If you proceed without, you are going to see processing errors.

PR#02: Is there a clear distinction between Males and females when it comes to the pattern of defaulting? Do one sex default more than the other? Produce a report that looks like this showing percent defaulted for both males and females.

```
+--------+-----+--------+-----------+
|SEX_NAME|Total|Defaults|PER_DEFAULT|
+--------+-----+--------+-----------+
|  Female|  591|   218.0|       37.0|
|    Male|  409|   185.0|       45.0|
+--------+-----+--------+-----------+
```

PR#03: How does marital status and level of education affect the level of defaulting? Does one category of customers default more than the other? Produce a report that looks like the following.

```
+---------+-----------+-----+--------+-----------+
|MARR_DESC|     ED_STR|Total|Defaults|PER_DEFAULT|
+---------+-----------+-----+--------+-----------+
|  Married|   Graduate|  268|    69.0|       26.0|
|  Married|High School|   55|    24.0|       44.0|
|  Married|     Others|    4|     2.0|       50.0|
|  Married| University|  243|    65.0|       27.0|
|   Others|   Graduate|    4|     4.0|      100.0|
|   Others|High School|    8|     6.0|       75.0|
|   Others| University|    7|     3.0|       43.0|
|   Single|   Graduate|  123|    71.0|       58.0|
|   Single|High School|   87|    52.0|       60.0|
|   Single|     Others|    3|     2.0|       67.0|
|   Single| University|  198|   105.0|       53.0|
+---------+-----------+-----+--------+-----------+
```

PR#04: Does the average payment delay for the previous 6 months provide any indication for the customer to default in the future? Produce a report that looks like the following.

```
+-----------+-----+--------+-----------+
|AVG_PAY_DUR|Total|Defaults|PER_DEFAULT|
+-----------+-----+--------+-----------+
|        0.0|  356|   141.0|       40.0|
|        1.0|  552|   218.0|       39.0|
|        2.0|   85|    41.0|       48.0|
|        3.0|    4|     2.0|       50.0|
|        4.0|    3|     1.0|       33.0|
+-----------+-----+--------+-----------+
```

PR#05: Come up with a prediction model that can predict whether the customer is going to default in the next month based on his/her history for the previous 6 months. Choose the best algorithm for this prediction model.

PR#06: The bank intends to group its customers into 4 groups based on their following attributes: SEX, EDUCATION, MARRIAGE, AGE_RANGE (ranges of 10). Come up with an algorithm to do this grouping based on their affinity to each other.