**Overview**

Banks and credit card clients often have a high risk in that they don't know if an indvidual will default on their credit card payment or not. The 10 features they have collected are:

* 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 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)

**IMPLEMENTATION**

1. **Loading and processing the data** : Do Data cleansing and making it to be enhanced to answer the required questions.
2. **Do the Analysis** : Perform the analysis using PYSPARK ,SPARKSQL

**ETL**: DATAFACTORY

**STORAGE**: AZURE BLOB, ADLS GEN2

**ANALYSIS**: AZURE SYNAPSE, AZURE DATABRICKS

**VISUALIZATION**: POWERBI

1.Load the file into a RDD

2.Remove header row

|  |
| --- |
|  |
|  | 3.Cleanup data. Remove lines that are not "CSV" |
|  | 4.Remove double quotes that are present in few records.  4.Convert into SQL Dataframe.  5.rounding of age to range of 10s. |

6.Normalize sex to only 1 and 2.

7.average billed Amount.

8.average pay amount

|  |
| --- |
|  |
|  | 9.Find average pay duration. |
|  | Make sure numbers are rounded and negative values are eliminated |

10.AVERAGE PERCENTAGE PAID

11.Add SEXNAME to the data using SQL Joins.

|  |
| --- |
|  |
| #Do analysis as required by the problem statement OF FIND CREDIT CARD DEFAULTERS |
| IMPLEMENT PYSPARK RDD PROCESS FOR BELOW SPARKSQL QUERIES | #Create a temp view |
|  | ccFinalDf.createOrReplaceTempView("CCDATA") |
|  |  |
|  |  |
|  | SpSession.sql("SELECT SEX\_NAME, count(\*) as Total, " + \ |
|  | " SUM(DEFAULTED) as Defaults, " + \ |
|  | " ROUND(SUM(DEFAULTED) \* 100 / count(\*)) as PER\_DEFAULT " + \ |
|  | "FROM CCDATA GROUP BY SEX\_NAME" ).show() |
|  |  |
|  |  |
|  | SpSession.sql("SELECT MARR\_DESC, ED\_STR, count(\*) as Total," + \ |
|  | " SUM(DEFAULTED) as Defaults, " + \ |
|  | " ROUND(SUM(DEFAULTED) \* 100 / count(\*)) as PER\_DEFAULT " + \ |
|  | "FROM CCDATA GROUP BY MARR\_DESC,ED\_STR " + \ |
|  | "ORDER BY 1,2").show() |
|  |  |
|  | SpSession.sql("SELECT AVG\_PAY\_DUR, count(\*) as Total, " + \ |
|  | " SUM(DEFAULTED) as Defaults, " + \ |
|  | " ROUND(SUM(DEFAULTED) \* 100 / count(\*)) as PER\_DEFAULT " + \ |
|  | "FROM CCDATA GROUP BY AVG\_PAY\_DUR ORDER BY 1" ).show() |
|  |  |