Please read the instructions carefully

Data is available in HDFS as '**Batch33_phdData.csv**' at the location '**/user/datasets/B33PHD**'

Data Description:

Below are the fields and their description

Field	Description
Target	**Person experienced Financial Distress in the past 2 years** **1 - Experienced Financial Distress** **0 - Not experienced any Financial Distress**
Utilization	Total balance on credit cards and personal lines of credit divided by the sum of credit limits
age	Age
FD_ind1	Number of times borrower has been in Financial distress for 30-59 days
Debt Ratio	Monthly debt payments
Monthly Income	Monthly income
FD_ind2	Number of Open loans
FD_ind3	Number of times borrower has been 90 days or more past due on repaying
FD_ind4	Number of mortgage and real estate loans including home equity lines of credit
FD_ind5	Number of times borrower has been 60-89 days past due but no worse in the last 2 years.
NumberOfDependents	Number of dependents in family excluding themselves

Missing values can be identified with the value, '**NA**' in the data.

Here our target feature name is '**Target**'.

Process this data and create machine learning models to predict if a new person is going to experience a financial distress for the next 2 years or not.

Activities:

Complete these activities using Spark.

You are free to use SparkML .

1. Read this data and create a data frame and verify the dataframe.

2. Display the count of rows and columns.

3. Give the percentage distribution of Target attribute and verify if it is a class imbalance problem or not.

4. After you create the dataframe in the first step, Target attribute will be in the first column of the dataframe, make it as the last column of the dataframe.

5. Find out which feature has how many numbers of missing values.

6. Fill the missing values for features as given below (**Do not delete the rows with Null/Missing values**):

- Utilization – Fill the average value of this feature for the records having null values for this feature
- Age - Fill the average value.
- FD_ind1 – Take the mode value of this feature and use that to fill null values.
- DebtRatio – Fill the average value.
- MonthlyIncome – Fill the average value.
- NumberOfDependents – Fill with Zero.
- For rest of the columns consider the mode of the respective feature, while filling out missing values

7. Once you fill out the missing values proceed for the steps to apply any Spark machine learning technique of your choice and try to see the accuracy of your models.

8. In this case we are trying to find out the person who will experience a financial distress so consider Target atrribute 1 as your positive case and 0 as negative.

9. Experiment with different machine learning models to maximize your recall, Try at least two different techniques and give the comparison of recall between them.

Note:

Export/Note all your pyspark commands into a text file and upload to piazza under CSE9099c module with the naming convention

B33PHD_<Enrollment_id>_<First_name>_<Last_name>_BigData_Script.txt

Ex.
B33PHD_1234_Abc_Xyz_BigData_Script.txt