Dear Participant,

Please find below the Assignment for Big Data on Spark course. Kindly make your submission before the deadline.

Telecom company is very concerned about its customers discontinuing with the service and opting to move to the one provided by the competitors. It is a major phenomenon is several service industries and is called ‘Churn’.  The company thinks that there are early indications available in the way a customer uses its service in predicting whether he/she is likely to churn.

**Data Set –**Churn.csv

The file Churn.csv contains some of the measures tracked by the company to see if it can predict the customers who are likely to churn and then take proactive action.

Your task is to perform exploratory data analysis on this data and ascertain if the data still has power to predict churn. If indeed such power exists, which variables have the capability to predict churn.

**Follow these steps to explore it --**

1. Read the data Churn.csv into pyspark.
2. Calculate summary statistics of variables.
3. Plot histogram of various variables.
4. Plot relationship of various variables with churn variable to explore if there is some interaction between them and churn.
5. Calculate correlation of the dependent variable with the independent variables.
6. What is your overall conclusion based on the analysis that you just did? ( Use Hypothesis Testing)

**Please Note:**

* Total marks allotted for assignment is 40.
* Please submit working code and output along with it in pdf/html/ipynb format.
* Please add necessary comments also in your solution file.
* Name format of files should always be **YourName\_Assignment\_BDS.pdf**

Marks Distribution is as follows --

|  |  |
| --- | --- |
| **Part** | **Marks** |
| Step - 1 | 6 |
| Step – 2 | 6 |
| Step – 3 | 6 |
| Step – 4 | 6 |
| Step – 5 | 6 |

|  |  |
| --- | --- |
| Step – 6 | 10 |