Project Proposal

Proposal:

In this project, I will like to solve some problems of financial sector with help of Spark, Python API, MLLib and ML API. The goal is to determine the credibility of customer for loan approval and help the bank managers to provide the loan to eligible person and increase the bank profits using Credit dataset

Description:

As financial companies embark on increasing their quality in services and acquire a better understanding of customers and their many preferences, the amount of data in augmenting day by day, collection happens frequently, and variety of data points is different and complex. So, in order to improvise efficiency of business in financial sectors, the ability of accessing, analyzing and managing vast and complex volumes of data which are evolving rapidly is a critical part for these companies. They want their data to be stored in their server in short time and in high amounts. Thus, Big Data comes into picture for this situation to help the financial and banking companies and respond to the requirements in a quick and efficient manner. One of the various problems of the banking sector is Risk and Capital Management. As there is increase in number of risk factors, the companies need a solution with higher computation power. Thus, the main purpose of the project is to provide with a model which helps in the identification of credibility of customer by building relations between different attributes of the customer's acquired assets and spending patterns. This will help the managers to gain a clear view of person's ability to pay back the loan in timely manner, minimize the losses of bank companies and increase the profits

Methodology:

When a bank gets a loan application, the bank must make a decision to approve the loan or not by factors of customer like account balance, purpose, account balance, credit amount, occupation and many such attributes. The risks included in bank decision are:

- If the customer is likely to repay the loan in timely manner, then not approving loan would be loss to bank
- If the customer is not likely to repay the loan in time, then approving loan for that person in not profitable

Thus, the project will help by giving an efficient predictive model to approve a loan to prospective applicant based on their attributes and guide the bank manager in making a profitable decision and less risk to banks.

The data will be loaded using Spark RDD (Resilient Distributed Datasets) and DataFrame, data preprocessing and tuning, data and attribute relation visualization using the matplotlib library of Python API and then deriving conclusion of credibility of customer. I will also try to explore different Machine Learning algorithms such as Linear Regression, Clustering and Naïve Bayes to compare them and get a better model for the project based on accuracy

Tools:

- Spark RDD and DataFrame
- Python API
- MLLib
- ML API
- Horton works Sandbox

References:

- https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)
- https://www.tutorialspoint.com/apache_spark/apache_spark_rdd.htm
- https://spark.apache.org/docs/latest/ml-guide.html
- https://www.tutorialspoint.com/pyspark/pyspark_mllib.htm