CS 724 - High Performance Computing and Big Data

Course Project Description

Herambeshwar Pendyala | 01130541

Title: Customer Churn Prediction

Dataset : Sparkify, a fictional Music Streaming Service similar to Spotify, by Udacity, http://udacity-dsnd.s3.amazonaws.com/

Project Objective

The primary objective of the project is to build a scalable model to predict the customer churn for a music streaming service, given features such as user activity on the website, events, user history and other user related data.

Project Description

Predicting customer churn is a common problem in customer facing business. The dataset contains user event logs of duration about 2 months. The log contains some basic information about the user as well as information about a single action. A user can contain many entries. This includes pages visited, timestamp, gender, location, artist, song details, playing next song, length of the song for every user. Using this data, we can predict whether the user is likely to stay or is more likely to churn. I intend to use pySpark, sparkSQL component to explore the dataset and understand the relation between different attributes of the dataset. Perform statistical analysis to get important features that can be used in predicting the customer churn. Finally, I would use these features to build a machine learning model to predict the customer churn and evaluate the model with necessary metrics.

As these predictions are largely data-driven and involve analyzing a huge number of user activity logs, we need a distributed way to handle a large dataset efficiently without having to fit it in our memory all at once. I want to use PySpark, as this allows me to work with RDDs which are capable of distributed processing and can achieve time efficiency. As the dataset is large, I would use a small subset of the data to explore and build a feasible machine learning model and determine its scalability by applying it to a large subset of data. This helps us to draw conclusions on which factors determine customer churn and get an in-depth understanding of spark.