Machine Learning Implementation

1. Data Set: This dataset describes a 5-star rating from Movie Lens, a movie recommendation service. It contains 100836 ratings and 3683 tag applications across 9742 movies. These data were created by 610 users
2. Collaborative Filtering: Collaborative filtering is a technique for creating automatic predictions (filtering) about a user's interests by collecting preferences or taste information from a large number of people. Alternating Least Square (ALS) is a matrix factorization algorithm, implemented in Apache Spark ML and built for large-scale Collaborative filtering problems. Its key parameters are:
   1. maxIter: the maximum number of iterations to run (set to 20)
   2. rank: the number of latent factors in the model (defaults to 10)
   3. regParam: the regularization parameter in ALS (set to 0.05)

The model was evaluated on 20% of total data and mse was found out to be 0.891.

1. Logistic Regression: PySpark Logistic Regression is a classification that predicts the dependency of data over each other in the PySpark ML model. It is a faster way of classification of data and works fine with larger data set with accurate results. Its key parameters are:
   1. maxIter: the maximum number of iterations to run (set to 20)
   2. regParam: the regularization parameter (set to 0.05)
   3. elasticNetParam: the combination of L1 and L2 regularizer (set to 0.8)

The modal was evaluated for both 80% training dataset and 20%test dataset and the rmse was found to be 3.6521 and 3.6585 respectively

<https://www.geeksforgeeks.org/find-duplicate-rows-in-a-dataframe-based-on-all-or-selected-columns/>

<https://www.kaggle.com/najeedosmani/movie-recommender-system-with-eda>

<https://towardsdatascience.com/machine-learning-with-pyspark-and-mllib-solving-a-binary-classification-problem-96396065d2aa>

<https://runawayhorse001.github.io/LearningApacheSpark/reg.html>

<https://www.kaggle.com/aigamer/movie-lens-dataset>

<https://www.kaggle.com/madhurisivalenka/basic-machine-learning-with-red-wine-quality-data>

<https://medium.com/@patelneha1495/recommendation-system-in-python-using-als-algorithm-and-apache-spark-27aca08eaab3>

<https://towardsdatascience.com/prototyping-a-recommender-system-step-by-step-part-2-alternating-least-square-als-matrix-4a76c58714a1>

<https://towardsdatascience.com/various-implementations-of-collaborative-filtering-100385c6dfe0>

<https://www.kaggle.com/eswarchandt/amazon-music-reviews?select=Musical_instruments_reviews.csv>