Pyspark - MLLib Part 1 - Regression

Agenda

This is the twelfth project in the Pyspark series. The eleventh project mainly focuses on how to optimize PySpark using Shared variables, Serialization, Parallelism and built-in functions of Spark SQL. This project mainly focuses on Regression in Spark MLlib. Apache Spark uses Spark MLlib to do machine learning. The main Machine Learning API for Spark is Spark.ml. For building ML pipelines, the package Spark.ml provides a higher-level API built on top of DataFrames. This project also includes implementation of Simple linear regression, Multiple linear regression and Random forest regression.

Tech stack:

➔Language: Python
➔Package: Pyspark
➔Services: Spark

Key Takeaways:

- Understanding the project overview
- Understanding the concept of Machine Learning
- Introduction to PySpark MLlib
- Understanding the Supervised learning
- Regression analysis
- Implementation of Simple linear regression
- Implementation of Multiple linear regression
- Implementation of Random forest regression
- Change datatype of column in PySpark
- Drop NA values in PySpark
- Data preprocessing in PySpark
- Splitting data into train data and test data in PySpark
- Validate the model in PySpark

Note:

- S3 link for dataset - s3://airlines123/airline/data.zip