IT 609 - Big Data Processing Lab 4

Teaching Assistant - Saran Pandian P

March 16, 2023

1 Aim

- learn how to use PySpark for applying Linear Regression.

2 Problem description

Spark provides APIs for various machine learning algorithms. In this lab assignment you build Linear regression model which predicts the selling price of a car based on a set of features like car name,year, km driven, number of seats, etc.

3 Implementation

3.1 Dataset

- You will use the Vehicle dataset from cardekho Car details v3.csv.

3.2 Exercise

RDD Tasks

- Read the dataset into a PySpark Dataframe.
- You are supposed to use all the columns except for torque (bonus task) for predicting the selling price using Linear regression.
- Use appropriate functions to transform the column data into numerical values and handle missing values.
- Compare the results obtained from standardization and without standardization.
- Split the data into train and test (80:20). Train the linear regression model using train data and evaluate the model on test data.
- Compute RMSE, MAE and R2 score for the test data.

4 Submission guidelines

You have to submit two files

- You have to submit your assignment in python notebook with proper comments and explanation. Print and show the results obtained in for every task in the notebook.

5 References

- PySpark SQL functions
- PySpark Linear Regression
- PySpark ML feature