Mason Harlan

12/7/2021

Project 2

GitHub Link: https://github.com/mgharlan/442_assignment2

Docker Hub Link: https://hub.docker.com/repository/docker/harlangmason/cs442

Brief Overview:

For a detailed overview of all the steps taken to develop this project check the README file in the git repository. I used jupyter notebook and pyspark with python to develop and train the models. The models implemented were linear regression, random forest regression, and random forest classification. Once the models were developed, I saved them into a git repo. From an EC2 instance, I installed docker and pulled the repo. I used a Dockerfile and the docker build command to build an image with my code from a pyspark jupyter image. This image has all the dependencies I needed and can be found at https://hub.docker.com/r/jupyter/pyspark-notebook. After the image was created, I pushed the image to my docker hub profile. To test that it worked I created a new EC2 instance, installed docker, pulled the image, and ran the image. The output of my models were printed as expected.