Kaggle Competitions with Spark

Miquel Vande Velde

December 9, 2022

1 Digit Recognition

1.1 Description

The kaggle competition entered was one to recognise hand written digits: `https://www.kaggle.com/competitions/digit-recognizer`. The most successfull machine learning models in this competition were convolutional neural networks (using the keras package). However, other classification algorithms seemed to be performing very good as well. This notebook achieved a (self reported) accuracy of 96% using a random forest classifier. Note that this accuracy was achieved on a test split of the training data, not the actual competition test data.

1.2 Comparing Spark with Notebook entries

1.2.1 Models

As mentioned above, the winning algorithm in Kaggle was, not surprisingly, a convolutional neural network, which is famous for image recognition. This model is not available in spark. However other classification techniques are also achieving high scores, with accuracies of more than 95%. As shown in the results below, I was able to reproduce these figures with spark. Using the default values offered in spark I only achieved an out of sample accuracy of 83%. Comparing the documentation of scikit and spark I noticed some differences in default configuration. In scikit learn the max depth of the trees is indefinite as opposed to 5 in spark and the amount of trees is 100, compared to 20 in spark. To find the best configuration I made use of spark's CrossValidator to tune these hyperparameters and reached an accuracy of 96%. The model achieved almost the same accuracy on submission see image below, indicating that it was not just overfitting.

1.2.2 Feature Engineering

In this competition feature engineering was not required to achieve high scoring results.

1.3 Results

Random Forest	
Algorithm	Accuracy
Kagel entry	96%
Spark default	83%
More trees	93%
More depth	96%

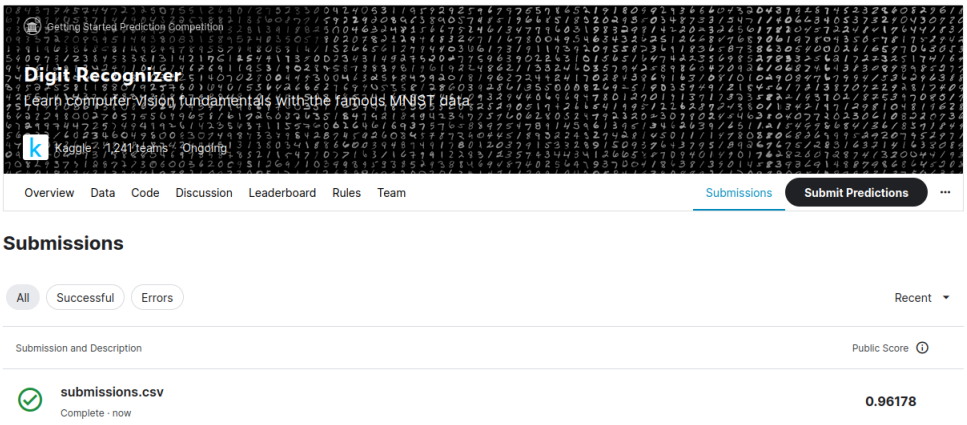

Figure 1: Random forest submission results.