Project Report

Team Members:

Ronit Sharma,

Utkarsh Pujari

GitHub Link: - https://github.ccs.neu.edu/utkarshpujari/SparkProject

Project Overview:

Our motivation for this project has been to learn how a machine learning model can be applied to data in a parallel environment. In the first task we have tried to run linear regression parallelly to predict house prices. We are working on data with around 80 features. As of now we have completed the linear regression part, we plan to increase the accuracy of our model by doing more research on feature extraction.

Input Data:

We are working with housing data, that has 80 features that describes the house, like LotArea, YearBuilt, etc., which we will use to predict the selling price of the house. We took the data from a Kaggle competition and it is already divided into training and testing data. The training data has 1461 rows and test data has 1460 rows to predict.

Task 1-Running Linear Regression and predicting house prices.

Overview

This task is focused on building a Machine Learning model (Linear Regression) in a parallel computing environment and predicting house process from that model.

Pseudo Code

```
● val inputDataRDDTest = test.csv
● val inputDataRDDTrain = train.csv
● val trainMap= inputDataRDDTrain.map(s=>parseRecordObject(s))
● val testMap= inputDataRDDTest.map(s=>parseRecordObject(s))
● val testDF = testMap.toDF
● val trainDF = trainMap.toDF
● val assembler = new Assembler().setInputCols(“Selected Features”).setOutputCols(features)
● val testDF2=assembler.transform(testDF1)
● val trainDF2=assembler.transform(testDF2)
```

- val lr = new LinearRegression()
- lr.fit(testDF2)
- lr.predict(trainDF2)

Algorithm and Program Analysis

We have used Linear Regression model as of now on training set. Out of 80 features we are using about 36 of those which are numeric and ran our prediction on test data. Few 'NA' values were converted to 0.

Our Plan is to convert non-numeric features into classes as well and extract those features which has more correlation to train our model. This will significantly improve our model.

Experiments

To start with we have trained our model based on all numeric features which are about 36 features on training data. There were many numeric columns with 'NA' as values, which for now we are converting into a 0. We have submitted our initial prediction at Kaggle, with a Root Mean Squared Logarithmic error as 0.44.

Speedup

Currently we have not run it on AWS. Locally it runs for 6secs with the current code.

Scaleup

Currently we have not run it on bigger dataset and AWS.

Result Sample

Predictions are sorted by id. These are the predictions.

```
[119983.98489053274]
[156626.08678829944]
[174710.8999169882]
[201434.72255561638]
[196320.55155298635]
[182283.4858034216]
[198298.52568115888]
[169742.7563206591]
[208876.54115646138]
[115883.96844387526]
[203689.49634292786]
[99331.08854661725]
[76336.97871382482]
```

Task2

We have not decided on task2.