Most Cost-Efficient and Accurate Way to Predict Rain in Australia

Nabeel Bacchus & Sammy Tawakkol

Introduction

- ML hypothesis: Can we predict if rain will occur the next day in Australia?
  - Dataset used: Rain in Australia
- We are splitting this dataset into three different sizes: 1000, 10000, and 100000.
  - Show scalability of each instance
- AWS EC2 Instances: t2.medium, t2.large, t2.xlarge
- Google Compute: e2.medium, e2.standard-2, e2-standard-4
- Measuring Performance:
  - Accuracy of Classification Model
  - Time to train Classification Model
  - Cost for time/Cost-efficiency

Changes Made

- Original project name: “AWS & Google Cloud Instance Performance for ML Processes”
  - Changed to “Most Cost-Efficient and Accurate Way to Predict Rain in Australia”
- Original instances were t2.small, t2.medium, and t2.large for Amazon, then e2.small, e2.medium, and e2.standard-2 for Google, but changed to equivalents of medium/large/xlarge due to lack of availability on Amazon’s end.

Results (Amazon, Time)

t2.xlarge	10^4	10^5	10^6
Regression	0.0094267	0.09503991	1.69566238
K-Near Neigh	0.0043462	0.01427754	0.16888775
Decision Tree	0.0225226	0.25350527	3.53805208

t2.large	10^4	10^5	10^6
Regression	0.0099416	0.09445295	1.61380679
K-Near Neigh	0.0047305	0.01442324	0.16761214
Decision Tree	0.0234650	0.23333830	3.67861899

t2.medium	10^4	10^5	10^6
Regression	0.0093694	0.09433571	1.68100277
K-Near Neigh	0.0043948	0.01457980	0.17124041
Decision Tree	0.0248510	0.24671627	3.86633959

Results (Google, Time)

e2.standard-4	10^4	10^5	10^6
Regression	0.0240739	0.1242056	0.4485327
K-Near Neigh	0.0047155	0.0142881	0.0439606
Decision Tree	0.0188957	0.2025898	0.7858159

e2.standard-2	10^4	10^5	10^6
Regression	0.0116520	0.1069996	0.7016204
K-Near Neigh	0.0039408	0.0168335	0.0646989
Decision Tree	0.0181566	0.2092190	1.0618188

e2.medium	10^4	10^5	10^6
Regression	0.0201533	0.1325864	2.5758995
K-Near Neigh	0.0052483	0.0171631	0.2944071
Decision Tree	0.0269514	0.2308174	3.1139359

Results (Amazon, Accuracy)

t2.xlarge	10^4	10^5	10^6
Regression	0.8263300	0.8466542	0.8500650
K-Near Neigh	0.7800555	0.7923977	0.8105066
Decision Tree	0.7812862	0.7747594	0.7925032

t2.large	10^4	10^5	10^6
Regression	0.8293188	0.8486936	0.8500650
K-Near Neigh	0.7813214	0.7933120	0.8105066
Decision Tree	0.7645486	0.7832202	0.7925032

t2.medium	10^4	10^5	10^6
Regression	0.8105066	0.8282288	0.8492563
K-Near Neigh	0.7817785	0.7936987	0.8105066
Decision Tree	0.7460177	0.7830795	0.7925053

Results (Google, Accuracy)

e2.standard-4	10^4	10^5	10^6
Regression	0.8230550	0.8375237	0.8413187
K-Near Neigh	0.7758538	0.7935246	0.7988614
Decision Tree	0.7561669	0.7667220	0.7858159

e2.standard-2	10^4	10^5	10^6
Regression	0.8200492	0.8378455	0.8386028
K-Near Neigh	0.7821847	0.7908936	0.8015903
Decision Tree	0.7739492	0.7747065	0.7775463

e2.medium	10^4	10^5	10^6
Regression	0.8185802	0.8401768	0.8422299
K-Near Neigh	0.7784665	0.7924826	0.8084335
Decision Tree	0.7754658	0.7759396	0.7856917

Analysis

- On average, Amazon is able to handle the smaller datasets and train them at a faster rate than Google. However, when we approach 10^6 size train/test, Google gets significantly faster.
- In terms of accuracy, Amazon, on average, is more accurate.
- In terms of cost, Amazon is a lot cheaper. Google adds up a lot per month in terms of prices.
- Google availability is difficult, whereas Amazon is easy to access across all instance locations

References

- Dataset: Rain in Australia: https://www.kaggle.com/jsphyg/weather-dataset-rattle-package

Thank You!