Comparing Scalable Predictive Analysis using Spark XGBoost Platforms

Savita Yadav, Samyuktha Muralidharan, Jongwook Woo*
Department of Information Systems, California State University
Los Angeles, California
[e-mail: {syadav5, smurali2, jwoo5}@calstatela.edu]
*Corresponding author: Jongwook Woo

Abstract

This paper compares the performance of scalable predictive analysis models using XGBoost in Big Data. The performance measurement is based on the training computing time and accuracy with AUR and Precision of a model. We developed XGBoost classification models with Airbnb listing dataset that predict the recommendation of the listings. The models are built in PySpark Rapids, BigDL, and H2O Sparkling with CPU and GPU on AWS EMR. We observed that BigDL with GPU is 25 – 50% faster training time than other platforms. H2O Sparkling has 5 - 7% better AUC and 0.7% better Precision than others.

Keywords: Airbnb Rating, Spark ML, Rapids, Big Data, Predictive Analysis, BigDL, H2O Sparkling Water

1. Introduction

XGBoost stands for Extreme Gradient Boosting, providing a machine learning library with distributed gradient-boosted decision tree boosting. It is famous as it has won many Kaggle competitions. We can build more accurate XGBoost models that work well with large-scale data sets.

Airbnb allows property owners to rent out their homes to people. It helps people to list, explore and book unique properties all over the world. In this paper, we developed models to predict the Airbnb Rating of the Listing that can classify if a property has a high or a low rating based on the features of the listing. The Airbnb Listings dataset is composed of 75 fields, which requires time-consuming data engineering for predictive analysis [3].

The Big Data cluster is a linearly scalable, distributed parallel computing system that stores and processes large-scale data sets [1]. We also add Intel BigDL and H2O Sparkling Water to the cluster, which provides machine learning libraries, including XGBoost.

2. Related Work

Savita et al. built and compared predictive models for Airbnb Rating using traditional and Big Data platforms such as Azure ML Studio and Spark ML. Their Random Forrest and GBT models show the highest Precision and AUC in Spark Big Data cluster. And, Multilayer Perceptron Classifier model predicts it with a shorter training time with a little less Precision and AUC [1].

Athor et al. predict Airbnb customers' behavior in classifying their reviews as binary-rated using attributes from the 66,630 reviews. They implement Tree Model, Random Forest Model, Least Absolute Shrinkage and Selection Operation and Logistic Regression Model, Artificial Neural Network, and Multi-Layer Perceptron in a traditional system [2]. We developed classification models for Airbnb Rating based on three different Big Data platforms running on AWS EMR Spark cluster. We compare the performance of the models with the Airbnb data set, where the models are built using XGBoost libraries as they are popularly scalable distributed parallel computing systems.

3. Cloud Computing Architecture

We have implemented our big data models using Hadoop, Hive, and Spark. As a Data Engineering phase, we transform the 1.9 GB dataset of Airbnb listings to the data frame of the categorical and numeric data. In the stage of Data Science, we set up BigDL and Sparkling Water platforms in addition to the EMR Spark that runs on the Spark Computing engine of the AWS Hadoop-Spark cluster. Then, we built predictive models to classify ratings in PySpark XGBoost libraries supported by each platform. Table 1 shows the software and hardware specification of the platforms.

Table 1. Technical Specifications

EMR Spark	BigDL	Sparkling
AWS EMR:emr-6.7.0, **Instance:** g4dn.xlarge, 2 x m3.xlarge, **Memory:** 15 GB, **GPU:** T4		
PySpark: 3.2.1	3.1.3	3.2.1
	BigDL: 2.0	**Sparkling:** 3.36

4. Background Work

We can use XGBoost Classification algorithms for large-scale datasets to build more accurate predictive models in rating prediction. Our models are to classify the listings as either high/low rated. The evaluation metrics are computing time to train models, AUC (Area Under Curve), Recall, and Precision.

4.1 PySpark Rapids in EMR

AWS EMR provides Spark ML as a Big Data solution for predictive analysis. We set up Rapids into EMR to build an XGBoost model with GPU and CPU. It produces much better accuracy and faster computing time by utilizing GPU cores in parallel [6].

4.2 Intel BigDL

Intel provides BigDL for distributed Deep Learning applications in the Spark cluster. One of its libraries, DLlib, is equivalent to Spark ML for Deep Learning and provides XGBoost [4].

4.3 H2O Sparkling Water

H2O presents Sparkling Water in Spark's ML algorithms with XGBoost library [5].

5. Experimental result

We develop three XGBoost classification models in PySpark Rapids, BigDL, and Sparkling on AWS EMR. The cluster consists of 3 nodes with an instance, g4dn.xlarge and two m3.xlarge. XGBoost classification models predict rating in Spark ML with 1.9 GB data. We set up the hyperparameters of the models as: {*learning_rate*: [0.1, 0.7], *max_depth*: [10, 15], *num_round*: 100, *nthread*: 1} where *learning_rate*: step size for trainin model, *max_depth*: maximum depth of a tree *num_round*: number of rounds for boosting, *nthread*: number of parallel threads to run XGBoost.

Review Scores Rating column of the data is a label that indicates the overall rating of a listing. We redefine the listings as high-rated with a value, 1 when **'Review Scores Rating'** >= 80. Otherwise, it is as low-rated with a value, 0. Thus, the **Review Scores Rating** column is transformed into a categorical column. The dataset is split into a 70:30 ratio for train and test data. We defined a pipeline for feature transformation and training the classifier model. The pipeline consists of a String Indexer, Vector Indexer, MinMax Scaler, Vector Assembler, and a two-class Classifier algorithm that trains a Binary Classification model. The pipeline preprocesses the data to remove outliers and handle null values.

Table 2. Feature Importance in XGBoost

Feature	Score
review_scores_value	0.839192276
review_scores_accuracy	0.065462551
review_scores_cleanliness	0.042763479

We calculate the feature importance score of the XGBoost in **Table 2**. It shows that the model regards "review scores value" as the essential feature when rating the listings.

Table 3. Evaluation metrics for Model Training Time

Platform	Processor	Train Time (sec)	Test Time (sec)
BigDL	GPU	24.2	0.6
H2O	GPU	32.6	0.32
H2O	CPU	34.3	0.25
EMR	GPU	36	0.4
BigDL	CPU	45	1.1
EMR	CPU	47.7	0.7

Table 3 and **Fig. 1** show the evaluation metrics of the models' training and testing times. The evaluation result shows that model training time in BigDL with GPU takes 24.2 sec, which is 25 – 50% faster than other platforms.

The short training time is essential, but we also take Precision and AUC as necessary for measuring accuracy. As we aim to reduce the number of False Positives, we consider the models that give a higher Precision value.

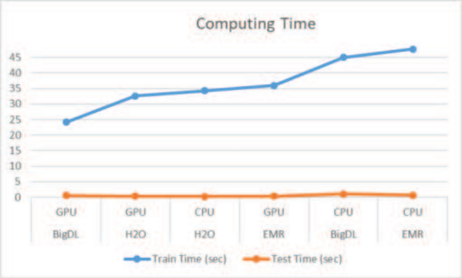

Fig. 1. Computing Time to Train Models

Table 4 shows that H2O Sparkling has 5 - 7% better AUC and 0.7% better Precision than other platforms no matter what it uses CPU or GPU. However, the AUCs of all models are close to 1.

Table 4. Models for Rating Prediction

Platform	Processor	AUC	Precision	Recall
H2O	CPU	0.9975	0.9830	0.9916
H2O	GPU	0.9975	0.9832	0.9915
BigDL	CPU	0.9518	0.9764	0.9767
BigDL	GPU	0.9483	0.9758	0.9761
EMR	GPU	0.9292	0.9771	0.9771
EMR	CPU	0.9248	0.9760	0.9758

6. Conclusions

The paper compares Big Data Spark platforms in XGBoost models for predicting Airbnb ratings. Rating prediction enables the customer to choose the relevant listings. It also helps the hosts know if their property is attractive compared to similar listings. Our Amazon Spark, BigDL, and H2O Sparkling platforms are composed of three nodes with CPU and GPU in AWS EMR and linealy scalable by adding more nodes, even with a hundred to thousand times more enormous data set.

The experimental result shows that BigDL with GPU is 25 – 50% faster for model training time than other platforms. It also shows that H2O Sparkling has 5 - 7% better AUC and 0.7% better Precision than others.

References

[1] S. Yadav, S. Muralidharan, S. Lee, J. Woo, “Scalable Predictive Analysis for Airbnb Listing Rating,” KSII The 13th International Conference on Internet (ICONI), Jeju Island, Korea, Dec 12-14 2021.

[2] Athor, S., & Marcel, C., “Rating prediction of peer-to-peer accommodation through attributes and topics from customer review,” Journal of Big Data, 8(1), 2021.

[3] Airbnb Ratings Dataset, 2021. Retrieved from https://www.kaggle.com/samyukthamurali/

[4] Intel BigDL XGBoost, 2022. Retrieved from https://github.com/intel-analytics/BigDL

[5] H2O Sparkling XGBoost, 2022. Retrieved from https://docs.h2o.ai/sparkling-water/3.2

[6] NVidia XGBoost, 2022. Retrieved from https://www.nvidia.com