# Comparing Regression Models Predicting the Price of Used Cars in Big Data

*Jo In Kang[a], Heta Parekh[b], Priya Ramdas[b], Seongwon Lee[c] and Jongwook Woo[b]
[a]Department of Applied Artificial Intelligence, Kwangwoon University, Korea
[b]Department of Information Systems, California State University, Los Angeles, USA
[c]School of Computer and Information Engineering, Kwangwoon University, Korea
[a,c]*{*kjikji956, swlee}@kw.ac.kr,* [b]*{hparekh2, pramdas2, jwoo5}@calstatela.edu*

2022 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia) | 978-1-6654-6434-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICCE-ASIA57006.2022.9954633

## Abstract

*We built and compared regression models to find the optimal solution using Spark Big Data engine, which predicts the price of used cars. We have the Cargurus dataset to perform predictive analysis, which is consumer data set as of 9.29 GB. The traditional systems have difficulty to handle the large scale data greater than hundreds of Mega-Bytes. Our Spark cluster can resolve this issue by storing and training the models with the large-scale data as distributed parallel computing systems. We adopt multiple regression algorithms to train models using Spark library in the big data environment. Then, we compare the models by measuring the computing time and accuracy. We observed that the GBT model shows the best accuracy in RMSE and R2 but relatively long computation time. The Random Forest model has a similar RMSE and R2 as GBT but with less computing time.*

**Keywords:** Big Data, Predictive Analysis, Machine Learning, Artificial Intelligence, Spark, Distributed Cloud Computing

## 1. Introduction

Since 2012, most of the consumers has purchased used cars, and many platforms offer excellent services to customers. The used car market is currently maintaining its growth in the states. As the market develops, the growth of the new car market is expected to slow down gradually. The demand for used car evaluation will continue to increase [1].

Cargurus is a well-known automotive shopping website headquartered in Cambridge, Massachusetts, where they are assisting millions of consumers with new cars and sellers' information. Also, dealers advertise on this site, and CarGurus connect with dealers and customers by displaying the inventory on the site. As one of good consumers data sets, the key factors present in the data are accident history, body type, price range, feature list, seller rating, etc.

Big data servers are being used in many areas. They stand out in areas that require many input data and require operation in real time [2]. The users can receive result of the model with only a simple specification device because model prediction performs operations on a huge server. In Big Data cluster with distributed parallel computing, calculations are performed in parallel on several servers, where it can store and train the models with the large dataset.

Big Data cluster has powerful computing resources and can handle large amounts of data. With this advantage, mobile device can collect the data in real time [3]. Using pre-trained deep learning models, Consumers enter the specific data at the tablet PCs or mobile electronics (low-power, low-sound devices) which connected with the cluster server. Consumers receive results in a short period of time using distributed processing cloud computing [4].

In this paper, we have used linear regression, random forest regression, Gradient boost tree, and recommendation models to find the predictive patterns in terms of the price, which would help the users to make informative decisions before customers purchase used cars from the CarGurus. To get accurate result in acceptable computing time, we used big data cluster and its machine learning technologies. We collect the used car dataset from Kaggle, and it consists of 3 million car records from 2005 to September 2020. The prediction models were built with the spark framework of Big Data cluster.

## 2. Related Work

There have been several research to predict used car prices. The artificial neural network ANN (Artificial Neural Network) and Machine Learning models have been used to predict the price [5]. Especially Among machine learning algorithms, the random forest regression model is showing relatively better performance in predicting used car prices [4, 6]. The models are implemented in the traditional systems that are not scalable for the growing data set.

The amount of data grows exponentially over time as a streaming data, distributed platforms are adopted to optimize the data processing [7]. It is to process and collect the large streaming data set but it is not to train the models to predict the prices.

We collect data sets and train the models using scalable big data systems, Spark. The spark computing engine runs on

Hadoop cluster and provides machine learning algorithms. The larger the data set, the more server we can add to maintain the performance. We trained and built models to predict the price of used cars. Then, we compare the models with the accuracy and computing time to find out which model is the best fit for the data set.

## 3. Specifications

**Table 1: Specification of Used Cars Dataset**

| | |
|---|---|
| File Size | 9.29 GB |
| File format | CSV |
| Years analyzed | 2006-2020 |
| Country | USA |
| No. of records | 3.0 million |

Table 1 is the detail of the dataset used to predict the prices. For prediction, data collected between 2006-2020 in the United States was used, and 9.29 GB of csv files were used as input. The computer resources are not good enough to build models in the traditional systems when the data size is greater than 500 Mega-Byte. Therefore, we use Big Data cluster to use distributed file systems and parallel computing engine.

**Table 2: Specification of Hadoop Cluster Sever**

| | |
|---|---|
| Cluster Version | EMR 6, Amazon Hadoop 3.2.1, Spark 3.1.1 |
| Number of Nodes | 5 |
| Memory size | 30 GB x 5 |
| CPU | 8 OCPUs x 5 |
| CPU Speed | 2.20 GHz |
| HDFS Storage capacity | 160 GB x 5 |

Table 2 is the specification of the cluster used in the experiment. We utilized the AWS (Amazon Web Services) cloud to launch EMR cluster consisting of five nodes. The Memory size is 30 GB, and the CPU has 8 cores with 2.20 GHz speed for each node (server).

## 4. Workflow

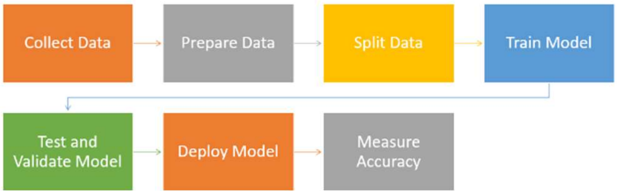

**Figure 2: Workflow Architecture**

Figure 1 shows the workflow to build the machine learning models. The first step is to collect the data to HDFS. The data preparation stage consisted of transforming the data into a structured format and converting it to desired data types, cleaning the missing values, removing duplicate data, and filtering unwanted rows and columns. Then, the third phase is to split the data into training and testing sets. The training set is to train and build a model. Training the models is one of the essential steps in any machine learning algorithm as it spends the most computing time while finding patterns of the data with its features and build a model with mathematical formula. The final step is to validate and test the model with accuracy. We implement models using regression algorithms so that the accuracy of the models can be measured through Root Mean Square Error (RMSE) and Coefficient of Determination (R2).

## 5. Methods and Experiments

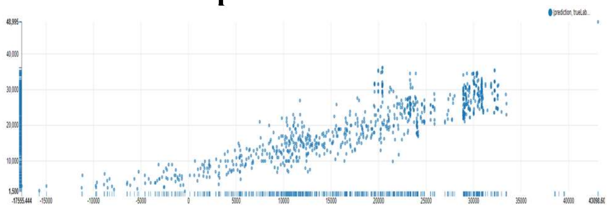

**Figure 2: The Prediction of Linear Regression Model**

Linear Regression (LR) is a model that combines a numerical set of input values to find the predicted outcome of the label. In Figure 2, the x-axis is the actual price of a used car, and the y-axis is the predicted value. We have feature columns to train the model. The features consist of all the columns of the car data, i.e., mileage, engine displacement, number of accidents, seller rating and horsepower. The Figure 2 shows the labels scattered majorly in the center giving an approximate diagonal line with $R^2$ of 0.735 and RMSE of 7766.5. When $R^2$ value is closer to 1, the better the model's accuracy is.

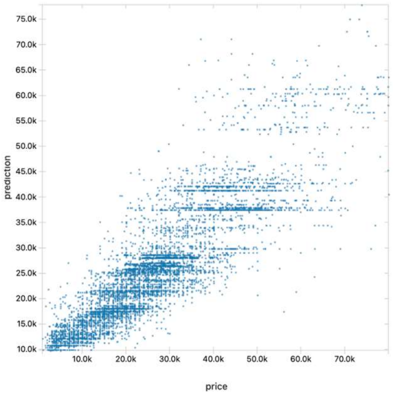

**Figure 3: The Prediction of Random Regression Model**

Random Forest Regression (RF) uses the ensemble learning method for regression. The ensemble learning method combines the predictions from multiple models with voting systems to make a more accurate prediction than a single model. In Figure 3, the x-axis is the actual price of a used car, and the y-axis is the predicted value with RMSE of 6319.9 and $R^2$ of 0.825. Compared to the LR model, this RF model's accuracy is better.

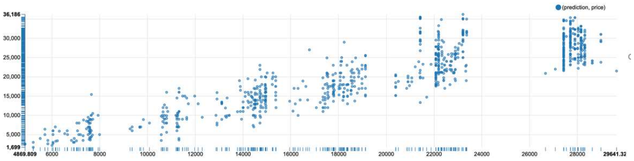

**Figure 4: The prediction of Gradient Boost Tree Model**

Gradient Boosted Decision Tree (GBT) is a forward learning ensemble method that obtains predictive results through incrementally improved estimation. The GBT method generalizes boosting to minimize the loss. In Figure 4, the x-axis is the actual price of a used car, and the y-axis is the predicted value. $R^2$ was 0.847 and the RMSE was 5912.7. This model produces better accuracy than RF, although the training time is longer than the LR and RF models.

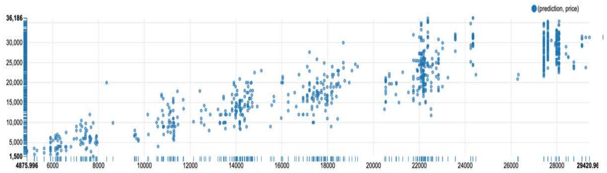

**Figure 5: The Prediction of Gradient Boost Tree Model with Cross Validation**

Figure 5 is the prediction of the GBT with the Cross Validation. The x-axis is the actual price of a used car, and the y-axis is the predicted value. Cross Validation build a model avoiding overfit and increases its generalization. Train Validation Split splits the input dataset into train and validation sets and uses evaluation metric on the validation set to select the best model. It has the similar generalization as Cross Validation with much faster computing time. $R^2$ for GTB with Cross Validation was 0.847 and with Train Validation Split was 0.848.

We built a model with the Factorized Machine Learning (FML) algorithm. It estimates interactions between features even in problems with huge sparsity. Thus, it fits with models for advertising and recommendation system. The RMSE and $R^2$ are 8910.7 and 0.651, respectively.

Table 3 shows the result of our experiments. We compared $R^2$ (R-squared values), RMSE (Root Mean Square Error), and Runtime. RMSE is a commonly used measure when dealing with the difference between the predicted value by the model and the actual value in the real environment. $R^2$ is a performance metric for regression analysis, mostly in linear algorithm. It is expressed as a value between 0 and 1, and it is judged that the closer it is to 1, the higher the accuracy. We can use $R^2$ to note if the models are good enough to practically use. Table 3 shows that GBT has the best accuracy with 10 ~ 66% smaller RMSE and 10 ~ 13% better $R^2$, and RF has the 10% less accuracy but with 19 ~ 88.7 % faster training time.

**Table 3: The Result of Experiment; $R^2$, RMSE and Runtime**

| Prediction model | $R^2$ | RMSE | Computing Time for Training (sec) |
|---|---|---|---|
| FML | 0.651 | 8,910.7 | 4875.29 |
| Linear Regression | 0.736 | 7,766.5 | 20.91 |
| Random forest | 0.825 | 6,,319.9 | 40.03 |
| GBT with CV | 0.847 | 5,913.6 | 355.01 |
| GBT | 0.847 | 5912.2 | 76.11 |
| GBT /w TV | 0.848 | 5,892.6 | 144.27 |

## 6. Conclusion

The market of used cars has been growing as the needs of consumers, and its data size has been getting bigger. Traditional systems have difficulty to train prediction models with a large-scale data because of the resource limitation. We show that Big Data systems, Spark cluster, can address this issue with distributed parallel computing. The experiment result and big data architecture will help the consumers.

We implement price prediction models in this market using Big Data cluster following the algorithms: LR, RF, GBT, and FML. It trains the models even though the data size is over 9 GB. We measure the computing time and accuracy of the models to find out the optimal model for predicting the price of the used cars. The GBT model has the best RMSE and R2 with 85% accuracy. For training time, GBT takes 76 secs to 355 secs. However, RF model has the similar accuracy and much faster training time, 40 secs.

## 7. Acknowledgements

We thank for this collaborative research Prof. Nanda Ganesan at the Department of Information Systems, and Prof. Tye Jackson, Dean of College of Business and Economics at

California State University Los Angeles. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2021R1F1A1060183) and by the MOTIE (Ministry of Trade, Industry, and Energy) in Korea, under the Fostering Global Talents for Innovative Growth Program (P0017308) supervised by the Korea Institute for Advancement of Technology (KIAT). This research was supported by a research credit from Oracle Cloud and Databricks University Alliance.

## References

[1] Yali Yang, Hao Chen, and Ruoping Zhang, “Suggestion on used Car Evaluation course teaching,” 2012 International Symposium on Information Technologies in Medicine and Education. IEEE, Aug-2012.

[2] Z.-X. Yang and M.-H. Zhu, “A Dynamic Prediction Model of Real-Time Link Travel Time Based on Traffic Big Data,” 2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS). IEEE, Jan-2019.

[3] H. Han, D. Wang, and J. Xu, “Classroom Assessment and Analysis System: Using Tablet PCs and Cloud Computing to Improve Effectiveness of Classroom Assessment,” 2013 International Conference on Advanced Cloud and Big Data. IEEE, Dec-2013.

[4] C.-S. Kim and S.-B. Son, “A Study on Big Data Cluster in Smart Factory using Raspberry-Pi,” 2018 IEEE International Conference on Big Data (Big Data). IEEE, Dec-2018.

[5] Varshitha, K. Jahnavi, and C. Lakshmi, “Prediction Of Used Car Prices Using Artificial Neural Networks And Machine Learning,” 2022 International Conference on Computer Communication and Informatics (ICCCI). IEEE, 25-Jan-2022.

[6] F. Wang, X. Zhang, and Q. Wang, “Prediction of Used Car Price Based on Supervised Learning Algorithm,” 2021 International Conference on Networking, Communications and Information Technology (NetCIT). IEEE, Dec-2021.

[7] A. Gupta and S. Jain, “Optimizing performance of Real-Time Big Data stateful streaming applications on Cloud,” 2022 IEEE International Conference on Big Data and Smart Computing (BigComp). IEEE, Jan-2022.