Random Forest Car-price Predict

Diego Miranda, Marcus Morris, Noah Gaffney

SUNY-New Paltz

Data Science - CPS 493 - 02

Professor Min Chen, Department of Computer Science

December 14, 2022

Task	Noah1	Marcus2	Diego	Total
Introduction	35%	35%	30%	100%
Background	45%	35%	20%	100%
Implementation	30%	45%	25%	100%
Experiment Results and Discussion	20%	20%	60%	100%
Conclusion	33%	33%	33%	100%
Other contribution and explain	33%	33%	33%	100%

Random Forest Car-price Predict

From the very start even before the group was whole the outset was to create a program capable of predicting the best price of a car. We decided that a random forest would be the best algorithm to give us that outcome. Because it's able to work in a trickle-down method it was able to take all the individual factors of the cars and create groupings that would accurately price a vehicle with a great level of specificity. There certainly were times at which we struggled to produce results, however, we are entirely delighted with the outcome of our labors.

Background

In 2021 nearly 15k cars were sold in the US, of those thousands around 82% paid over sticker prices for their new vehicles. The problem is that buyers don't have the tools to assess the value of their prospective vehicles. If there was a tool that could compute the value of a car or other aspects provided what results might we see? With a tool that would allow us to assess the value of the car, it's reasonable to say that around 24 million dollars of savings could be provided for buyers within a year of a tool of the like becoming available.

Approach and Implementation

We used Mapreduce to count the number of makes, bodies, horsepower of each car, mile per gallon, and price. We then implemented Random Forest to accurately predict prices. We decided on random forest being the best algorithm for us to work with as we are working on multiple decision trees there is a chance of less over-fitting. Additionally, since it runs on a larger data set, the accuracy is higher, and you can estimate the missing values for the same reason. Because of the efficiency and preciseness of the random forest, we were able to produce results of nearly 92% accuracy. Predominant research was performed through youtube tutorials. Everything besides MapReduce was utterly new to us.

Experiment Result/Discussion

When implementing the Random Forest Algorithm, we ran into issues. While the algorithm itself was straightforward with the help of jupyter notebook and pandas, the implementation of Hadoop with the code led to issues. Our first issue was that we had Hadoop with eclipse. When doing our word count assignment, we were able to set up Hadoop problem-free, given that we were provided with a tutorial for Windows OS. Upon looking at what we did, we then figured we could use this same principle, or at least the Map and Reduce functions for our code. The thing was that even though we were able to use the Map and Reduce for the dataset we chose, first by converting it into a .txt file, we then had no idea how to implement Hadoop into our algorithm in a way that would make sense for us. So we chose to focus on the algorithm and Hadoop as 2 separate things.

So for the Random Forest Algorithm, we made sure to focus on the attribute of “price” as this would be the main factor for the goal of our project, being to predict car prices using random forest. After doing all the necessary prerequisites, such as converting all the data types of integer to float so it could be read into our algorithm and used for training and testing sets, we were able to predict the price. However, when comparing the actual values to predicted values, we found that he had a pretty large margin of error. Despite having an Accuracy of 91.82%, our Mean Absolute Error, Mean Squared Error, and Root Mean Squared Error was found to have super high values. Normally, we would want these values to be 0, or at least as close as possible. However, it was found that they were well into the thousands.

However, despite having some errors with our specific implementation of the Random Forest Algorithm, we were able to get some impressive results. We believe that if we had better training and testing set using more specific attributes, and/or implementing Hadoop better, we would've been able to achieve better results, or at least more highly accurate results, possibly with 95%+ accuracy.

Conclusion

We expected this project to be more straightforward and streamlined when it came to Mapreduce linking with random forests. Although Random Forest is able to handle more variables to give a more accurate prediction, our implementation with chosen attributes for the training and testing sets, and/or lack of being able to implement Hadoop into our code gave us less than desirable results. With this, we cover lots of things from visualizing data, cleaning data, evaluating models, and doing the prediction. The accuracy of the model might not be so high but we get to see the process of implementing random forest predictions and can see that with the right use of attributes, and training/testing sets, this algorithm is more than capable to producing future companies and automotive buyers with accurate results based on attributes of cars they look at.

References

Alsenani, D. (2020, April 22). US cars dataset. Kaggle. Retrieved December 14, 2022, from https://www.kaggle.com/datasets/doaaalsenani/usa-cers-dataset

Carlier, M. (2022, April 25). U.S.: Average selling price of new vehicles 2021. Statista. Retrieved December 14, 2022, from https://www.statista.com/statistics/274927/new-vehicle-average-selling-price-in-the-united-states/#:~:text=In%20the%20United%20States%2C%20the,in%202021%20than%20in%202020.

Darlington, A. (2017, September 1). Car Evaluation Data Set. Kaggle. Retrieved December 14, 2022, from https://www.kaggle.com/datasets/elikplim/car-evaluation-data-set

Gokce, E. (2020, January 10). Predicting used car prices with machine learning techniques. Medium. Retrieved December 14, 2022, from https://towardsdatascience.com/predicting-used-car-prices-with-machine-learning-techniques-8a9d8313952

Machine learning random forest algorithm - javatpoint. www.javatpoint.com. (n.d.). Retrieved December 14, 2022, from https://www.javatpoint.com/machine-learning-random-forest-algorithm

MapReduce algorithm. TutorialsCampus. (n.d.). Retrieved December 14, 2022, from https://www.tutorialscampus.com/map-reduce/algorithm.htm#:~:text=MapReduce%20is%20a%20Distributed%20Data,efficient%20way%20in%20cluster%20environments.

O'Brien, S. (2022, February 17). With limited inventory due to a computer chip shortage, 82% of consumers are paying above sticker price for a new car. CNBC. Retrieved December 14, 2022, from https://www.cnbc.com/2022/02/17/more-than-80percent-of-consumers-are-paying-above-sticker-price-for-new-car.html#:~:text=To%20that%20point%2C%2082%25%20are,and%200.3%25%20in%20early%202020.

R, S. E. (2022, November 30). Random Forest: Introduction to random forest algorithm. Analytics Vidhya. Retrieved December 14, 2022, from https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

Reese, A. (2021, May 6). Used cars dataset. Kaggle. Retrieved December 14, 2022, from https://www.kaggle.com/datasets/austinreese/craigslist-carstrucks-data

Wu, Q., Wang, H., Yan, X., &amp; Liu, X. (2018, November 26). MapReduce-based adaptive random forest algorithm for multi-label classification - neural computing and applications. SpringerLink. Retrieved December 14, 2022, from https://link.springer.com/article/10.1007/s00521-018-3900-8

Yiu, T. (2021, September 29). Understanding random forest. Medium. Retrieved December 14, 2022, from https://towardsdatascience.com/understanding-random-forest-58381e0602d2