# Enhancing Carpooling through XG-Boost & Random Forest Regressor

# 1st Mr. Vaibhav Rautela

*Department of Computer Science and Engineering, D*ehradun, India

1[VAIBHAVRAUTELA.21011354@gehu.ac.in](mailto:mailtoVAIBHAVRAUTELA.21011354@gehu.ac.in)

***Abstract — India is a huge country with the highest population of one billion four hundred twenty-eight million six hundred twenty-seven thousand six hundred sixty-three Cities as the population is increasing day my day.*** [***Non-renewable energy resource***](https://www.bing.com/ck/a?!&&p=b4748578036f0b34JmltdHM9MTcxMzIyNTYwMCZpZ3VpZD0zODBkMmQ0YS03ZDgwLTZiYmMtMDQ4MS0zOTRjN2M3MjZhMGEmaW5zaWQ9NTgyMQ&ptn=3&ver=2&hsh=3&fclid=380d2d4a-7d80-6bbc-0481-394c7c726a0a&psq=oil+is+renewable&u=a1aHR0cHM6Ly93d3cubGlucXVpcC5jb20vYmxvZy9pcy1vaWwtcmVuZXdhYmxlLW9yLW5vbnJlbmV3YWJsZS8&ntb=1) ***consumption by human is also increase (for example: Oil****,****Uranium****,****Natural Gas, Coal)****.* ***Due to the abrupt shift in population, traffic congestion, and other issues, the expansion that has altered many cities has resulted in extremely long commutes and altered travel times, which is altering the public transportation systems. These reasons encourage people to use private automobiles, which increases traffic, increases idle hours, the use of a vehicle's its limit is reduced, raises the cost of transportation, and greatly raises vehicle emissions—one of the main environmental issues of our day. Carpooling lowers each person's travel expenses, such as fuel prices, taxes, plus the strain of traveling, by utilizing one vehicle for many users. Algorithms that are used for processing the datasets: The XGBoost model, Vector Autoregression, Support Vector Regression, K-Nearest Neighbors.***

*Keywords— eXtreme Gradient Boosting (XGBoost), Vector Autoregression (VAR), Support Vector Regression (SVR), K-Nearest Neighbors (KNN), Parking-Solution.*

I. INTRODUCTION:

Taking use of the empty seats in personal transportation vehicles might lower the costs associated with single occupancy cars—both financial and environmental. Sharing a vehicle and sharing a ride aim to fill those vacant seats by removing extra cars from the road, which lowers pollution and traffic, and by fostering social contact. Nevertheless, carpool booking in the past frequently restricted users to preset rider groups and constant schedules—carpooling to an identical location at the same moment with a certain individual or group of individuals. Carpooling arrangements can be established using a variety of platforms, such as open websites, social networks, markets, company websites, apps for smartphones, ridesharing companies, and drop-off, places. It makes it possible for people to travel together to comparable starting and ending points. The study of mathematical algorithms that can automatically get better with practice and with the use of information is known as machine learning, or ML. It also makes it much easier for anybody to share a ride and can be used to anticipate many characteristics that are needed to execute normal carpooling and long-distance carpooling concurrently [1].

II. LITERATURE REVIEW

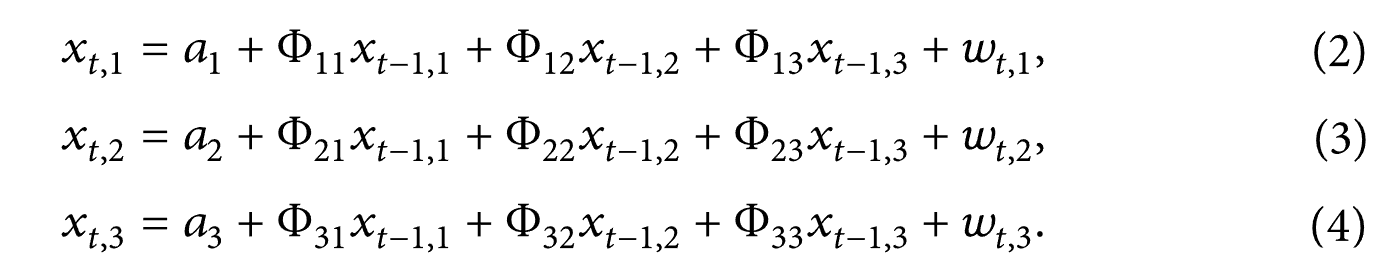
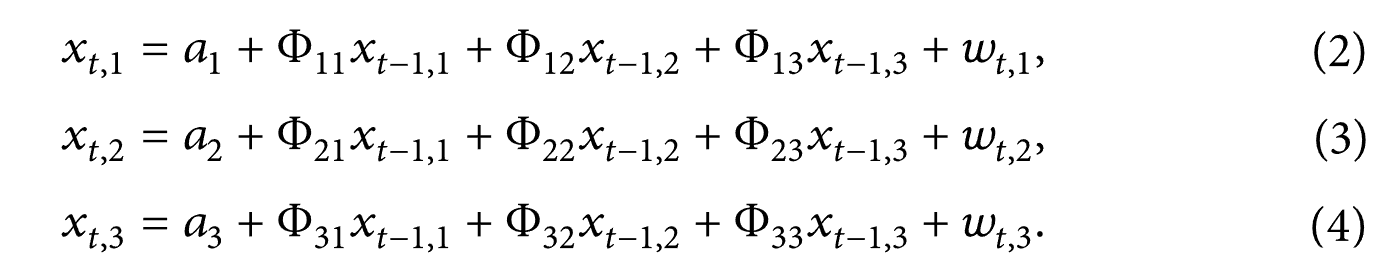
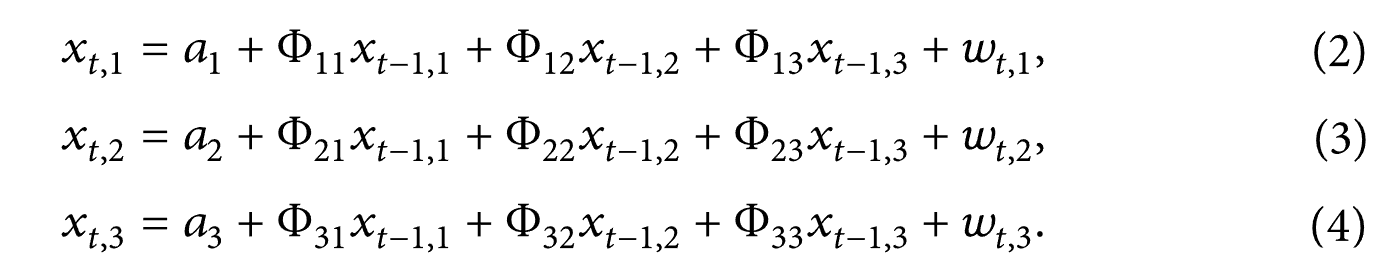
Deep Reinforcement Learning algorithm that was used for ride sharing by [Abubakr O. Al-Abbasi](https://ieeexplore.ieee.org/author/37085799624) [Arnob Ghosh](https://ieeexplore.ieee.org/author/37086049718) and [Vaneet Aggarwal](https://ieeexplore.ieee.org/author/37302263100) published on [12](https://ieeexplore.ieee.org/xpl/tocresult.jsp?isnumber=8939300&punumber=6979), December 2019 [2] write a research paper on the car pooling using the deep reinforcement learning.[3] Subhieh El Salhi1 , Fairouz Farouq2 , Randa Obeidallah3 , Yousef Kilani4 , Esra’a Al Shdaifat car-pooling system that was based on the KNN algorithem A Case Study in Hashemite University 2019 [4] [Marina Haliem](javascript:void(0);), [Vaneet Aggarwal](javascript:void(0);), [Bharat Bhargava](javascript:void(0);) wrote a research paper on dispatching using deep reinforcement learning. In 2020 [5] [Alejandro Lugo](https://ieeexplore.ieee.org/author/163071411440132); [Nathalie Aquino](https://ieeexplore.ieee.org/author/37085864770); [Magalí González](https://ieeexplore.ieee.org/author/37085870597); [Luca Cernuzzi](https://ieeexplore.ieee.org/author/37669798300); [Ronald Chenú-Abente](https://ieeexplore.ieee.org/author/37085635707) go with the Decongesting traffic through carpooling using automatic pairings in 2020.[6] [Junchao Lv](https://link.springer.com/chapter/10.1007/978-3-030-82136-4_55#auth-Junchao-Lv), [Linjiang Zheng](https://link.springer.com/chapter/10.1007/978-3-030-82136-4_55#auth-Linjiang-Zheng), [Longquan Liao](https://link.springer.com/chapter/10.1007/978-3-030-82136-4_55#auth-Longquan-Liao) & [Xin Chen](https://link.springer.com/chapter/10.1007/978-3-030-82136-4_55#auth-Xin-Chen) Ride-Sharing Matching of Commuting Private Car Using Reinforcement Learning 07 August 2021. [7] Kaushik Manchella; Abhishek K. Umrawal; Vaneet Aggarwal FlexPool: A Distributed Model-Free Deep Reinforcement Learning Algorithm for Joint Passengers and Goods Transportation. On 15 January 2021. [8] Manish Kumar Pandey, Anu Saini, Karthikeyan Subbiah, Nalini Chintalapudi andGopi Battineni 5, Improved Carpooling Experience through Improved GPS Trajectory Classification Using Machine Learning Algorithms on **3** August 2022. [9] Hoseb Abkarian, Ying Chen, and Hani S. Mahmassani on Understanding Ridesplitting Behavior with Interpretable Machine Learning Models Using Chicago Transportation Network Company Data on 2022. [10] Mohd Anas; Gunavathi C; Kirubasri G on Machine Learning Based Personality Classification for Carpooling Application on 2023 that was based on NLP

III. METHODOLOGY

1. *First Algorithm: Vector Autoregression (VAR):*

When more than two or more time series affect one another, a forecasting technique called vector autoregression is employed. Because the predictors represent historical lags of the model as well as lags of the series, it is referred to as an autoregressive model. Assume that all three time series factors that we are measuring are represented as, .

The following is the vector autoregression structure for class 1, or regression analysis (1), as stated:

(1)(2)  (3)

The variable acts as the model's endpoint and is a k-vector of constants. is a k-vector of error terms and is a time-dependent (k × k)-matrix.

1. *Second Algorithm: eXtreme Gradient Boosting (XGBoost)*

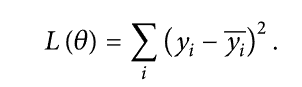
When three or additional time periods affect one another, a method for forecasting called vector autoregression is employed. Since the variables that predict represent historical lags of the model as well as lags of the series, it is referred to as a model with autoregressive properties. Assume that the three period variables that we are measuring are represented as . The following is the scalar autoregression framework for level 1, or a VAR (1), as stated:

1. *Third Algorithm: eXtreme Gradient Boosting (XGBoost)*

Friedman et al.'s, for example and effective and scalable gradient that improves technique has been included by XGBoost. A tree-based method for learning and a powerful linear regression solution are included in the package. XGBoost minimizes the departure despite adding the most recent forecast after fitting the new model to the new residuals of the prior prediction. Its use of "a more established system institutionalization for managing excessive fitting, resulting in its superior effectiveness" separates it from different models. —Tianqi Chen. When dealing with supervised tasks involving learning, such as predicting an objective variable using training data xi, XGBoost is employed. Following the selection of the goal variable, we must establish the goal function, which is comprised of the term for regulation and the training loss and is used to assess how well the model matches the trained data:

 (4)

where Ω is the normalization period of time, L is the learning function for loss, and θ is the parameters that we have to extract from the information we have. A frequently used value for L is the mean-squared mistakes, as provided by:

 (5)

1. *Fourth Algorithm: Support Vector Regression (SVR)*

Vapnik and Chervonenkis created the groundwork for support vector machines (SVM), and the approach is becoming more and more well-liked. Support vector classification (SVC) refers to the SVM foundations that handle classification issues, while support vector regression (SVR) refers to the SVM principles that deal with prediction and forecasting.

The following common kernels are employed in SVR modeling:

(1) Linear kernel: x ∗ y  
(2) [(x ∗ xi) + 1d is the polynomial kernel  
(3) RBF (radial basis function): exp {-ϒ |x − xi|2}

1. *Fifth Algorithm*: K-Nearest Neighbors (kNN):

K-nearest neighbors (kNN) is a popular approach for pattern recognition classification that is both effective and straightforward. Since it is a difference-based algorithm, it is implied that it makes the implicit assumption that a pair of points would be closer in appearance if their distance were reduced. Compared to KNN regression, the KNN classification technique is significantly more widely used. How near each training point Xi is to the testing point Xt is determined using the Euclidean distance.

1. Sixth Algorithem: Random Forest Regressor

The Random Forest Regressor algorithm is a popular choice for regression tasks due to its ensemble learning approach, which combines the predictions of multiple decision trees. Each decision tree is trained on a random subset of the data, and a random subset of features is considered at each node, reducing overfitting and improving generalization. By employing bagging and random feature selection techniques, Random Forest Regressor enhances the diversity among the trees, leading to more robust predictions. Hyperparameters such as the number of trees, maximum tree depth, and minimum samples per split can be tuned to optimize performance. Additionally, Random Forest Regressor provides insights into feature importance, aiding in feature selection and understanding the dataset's underlying relationships. With its scalability and efficiency in handling large datasets, Random Forest Regressor is widely used across various domains for accurate regression predictions.

##### Prediction Result:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | MAE | | MSE | RMSE | MAPE | | RMSLE | | | |
| VAR | 1.20552 | | 1.06025 | 1.4851 | 1.15297 | | 0.98598 | | | |
| XGBoost | **0.00975** | | **0.07272** | **0.09868** | **0.07189** | | **0.30398** | | | |
| Vector Autoregression | 2.77452 | | 1.19834 | 1.77467 | 1.45087 | | 1.91974 | | | |
| Support Vector | 1.86949 | | 1.15297 | 1.69395 | 1.3575 | | 1.7977 | | | |
|  |  | Fig 1: Machine learning for univariate time series evaluation results. | | | |  | |  |  |

With respect to VAR, SVM, and KNN, XGBoost decreased the error by 93.14%, 99.14%, 93.36%, 69.17%, 93.76%; 93.69%, 99.48%, 94.17%, 83.09%, 94.70%); and 93.93%, 99.65%, 94.44%, 84.17%, and 95.05%.

|  |  |
| --- | --- |
| Algo | Time |
| *eXtreme Boosting* | 10.2203 |
| Vector Autoregression | 34.5798 |
| KNN | 20.4125 |
| Support Vector | 65.2421 |

RESULT:

XGBoost yielded the best results, next to VAR, SVR, and KNN, according to the examination of the assessment metrics and the outcomes produced with the machine learning models. Full feature extraction, a strong fitting effect, and excellent prediction accuracy were only a few of the XGBoost model's many predictive benefits. XGBoost yielded the best results, next to VAR, SVR, and KNN, according to the examination of the assessment metrics and the outcomes produced with the machine learning models. Full feature extraction, a strong fitting effect, and excellent prediction accuracy were only a few of the XGBoost model's many predictive benefits.   
Secondly, unpredictable and irregular patterns were not captured by the SVR prediction series. As a result, it failed to perform effectively; in contrast, random stroll patterns were recorded by the XGBoost and VAR prediction series. Third, because of the large number of inputs, KNN fared the worst when compared to the other machine learning models.

CONCLUSION

By utilizing several machine learning models on time series with multiple variables, this research article attempts to forecast automobile usage and explores the elements that contribute to increased accuracy in predicting. When comparing the various deep learning and machine learning models using MAE, MSE, RMSE, MAPE, and RMSLE, it can be shown that the hybrid model produces significantly less errors than the solo models. The findings of the experiment indicate that the using of hybrid model provide more accuracy and make our model more perfect.

REFERENCES:

1. A. Agrawal et al., “VQA: visual question answering,” Aug. 15, 2020.
2. Reinforcement Learning algorithm that was used for ride sharing by “[Abubakr O. Al-Abbasi](https://ieeexplore.ieee.org/author/37085799624)”, “[Arnob Ghosh](https://ieeexplore.ieee.org/author/37086049718)” and “[Vaneet Aggarwal](https://ieeexplore.ieee.org/author/37302263100)” published on [[12](https://ieeexplore.ieee.org/xpl/tocresult.jsp?isnumber=8939300&punumber=6979), December 2019]
3. “Subhieh El Salhi”, “Fairouz Farouq”, “Randa Obeidallah”, “Yousef Kilani”, “Esra’a Al Shdaifat” car-pooling system that was based on the KNN algorithem, A Case Study in Hashemite University [2019]
4. “[Marina Haliem](javascript:void(0);)”, “[Vaneet Aggarwal](javascript:void(0);)”, “[Bharat Bhargava](javascript:void(0);)” wrote a research paper on dispatching using deep reinforcement learning. [2020]
5. “[Alejandro Lugo](https://ieeexplore.ieee.org/author/163071411440132)”, “[Nathalie Aquino](https://ieeexplore.ieee.org/author/37085864770)”, “[Magalí González](https://ieeexplore.ieee.org/author/37085870597)”, “[Luca Cernuzzi](https://ieeexplore.ieee.org/author/37669798300)”, “[Ronald Chenú-Abente](https://ieeexplore.ieee.org/author/37085635707)” go with the Decongesting traffic through carpooling using automatic pairings in [2020].
6. “[Junchao Lv](https://link.springer.com/chapter/10.1007/978-3-030-82136-4_55#auth-Junchao-Lv)”, “[Linjiang Zheng](https://link.springer.com/chapter/10.1007/978-3-030-82136-4_55#auth-Linjiang-Zheng)”, “[Longquan Liao](https://link.springer.com/chapter/10.1007/978-3-030-82136-4_55#auth-Longquan-Liao)” & “[Xin Chen](https://link.springer.com/chapter/10.1007/978-3-030-82136-4_55#auth-Xin-Chen)” Ride-Sharing Matching of Commuting Private Car Using Reinforcement Learning [07 August 2021].
7. “Kaushik Manchella”, “Abhishek K. Umrawal”, “Vaneet Aggarwal” FlexPool: A Distributed Model-Free Deep Reinforcement Learning Algorithm for Joint Passengers and Goods Transportation. [15 January 2021].
8. “Manish Kumar Pandey”, “Anu Saini”, “Karthikeyan Subbiah”, “Nalini Chintalapudi” and “Gopi Battineni”, Improved Carpooling Experience through Improved GPS Trajectory Classification Using Machine Learning Algorithms on [**3** August 2022].
9. “Hoseb Abkarian”, “Ying Chen”, and “Hani S. Mahmassani” on Understanding Ridesplitting Behavior with Interpretable Machine Learning Models Using Chicago Transportation Network Company Data on [2022].
10. “Mohd Anas”, “Gunavathi C”, “Kirubasri G” on Machine Learning Based Personality Classification for Carpooling Application on [2023] that was based on NL.