Prediction of Fuel Efficiency using Machine Learning

Abstract

The automotive industry is extremely competitive. With increasing fuel prices and picky consumers, automobile makers are constantly optimizing their processes to increase fuel efficiency. But what if one could have a reliable estimator for a car's mpg given some known specifications about the vehicle? Then, one could beat a competitor to market by both having a more desirable vehicle that is also more efficient, reducing wasted R&D costs and gaining large chunks of the market. Using machine learning on the UCI Machine Learning Repository: Auto MPG dataset, which machine learning methods are most successful, how data collection affects the prediction and which features are most influential for fuel consumption are evaluated. The multivariate dataset consists of 3 multivalued discrete and 5 continuous attributes in addition to 2 attributes which are derived from the dataset. Acceleration on power proved to be the best estimator among the attributes. Regression is applied on the dataset where all the models could predict fuel consumption accurately. Various Machine Learning models have an absolute relative error less than 10%. The random forest model is proved to have the highest accuracy and runs faster, making it suitable for wide application. This method lays a foundation for monitoring database improvement and fine management of urban – rural transportation fuel consumption.

1. Introduction

Vehicle energy consumption and pollutant emissions are key problems for the healthy and sustainable development of urban - rural transportation. With the continuous growth of car ownership, the energy consumption of private cars increased 5.1 times, from 130.12 to 680.34 million gallons of fuel, from 2005 to 2015. Based on growth of the population, GDP, and the proportion of secondary and tertiary industries, the trend of future transportation energy consumption can be predicted. The energy consumption of private cars will continue to increase before 2023, when it is expected to reach 1170.38 million gallons of fuel. Therefore, reducing energy consumption has become one of the most important challenges in the transportation field.

This study evaluates methods of machine learning (ML) with help from statistical analysis for predicting fuel consumption in vehicles. The idea is to use historical data describing primary and secondary attributes of a vehicle to predict fuel consumption in gallons per mile. The general problem description is to examine a large number of attributes describing fuel consumption and to employ ML methods to find a regression from such attributes to predict fuel efficiency.

Fuel consumption models for vehicles are of interest to manufacturers, regulators, and consumers. They are needed across all the phases of the vehicle life-cycle. This paper, is focused on modeling average fuel consumption for vehicles. In general, techniques used to develop models for fuel consumption fall under three main categories:

- Physics-based models, which are derived from an in depth understanding of the physical system. These models describe the dynamics of the components of the vehicle at each time stamp using detailed mathematical equations.

- Machine learning models, which are data-driven and represents an abstract mapping from an input space consisting of a selected set of predictors to an output space that represents the target output, in this case average fuel consumption.

- Statistical models, which are also data-driven and establish a mapping between the probability distribution of a selected set of predictors and the target outcome.

Trade-offs among the above techniques are primarily with respect to cost and accuracy as per the requirements of the intended application. The studies of the past are mainly based on 11 features dataset. This dataset is a slightly modified version of the dataset provided in the StatLib library. In line with the use by Ross Quinlan (1993) in predicting the attribute "mpg", 8 of the original instances were removed because they had unknown values for the "mpg" attribute. The data concerns city-cycle fuel consumption in miles per gallon, to be predicted in terms of 3 multivalued discrete and 5 continuous attributes (Quinlan, 1993).

Various Machine Learning models employed have an absolute relative error less than 10%. The random forest model is proved to have the highest accuracy and runs faster, making it suitable for wide application. The work done by L. Breiman, by applying “Random forests,” Machine Learning, plays along the notion established in the paper. Moreover the work done by H. Drucker, J. C. Chris, L. Kaufman, A. Smola, and V. Vapnik, “Support vector regression machines,” in Advances in Neural Information Processing Systems gives an overview of the SVM mechanism but has a relative error more than the Random Forest model. D. Yang, M. Li, and X. Ban, “Real-time on-board monitoring method of gasoline vehicle fuel consumption based on OBD system reflects on the statistical modelling while works done by X.-h. Zhao, Y. Yao, Y.-p. Wu, C. Chen, and J. Rong, on “Prediction model of driving energy consumption based on PCA and BP network and works of W. J. Zhang, S. X. Yu, Y. F. Peng, Z. J. Cheng, and C. Wang, on “Driving habits analysis on vehicle data using error back-propagation neural network algorithm employs Deep Learning and Neural Network Architecture to produce the results.

Air pollution is one of the world’s single biggest environmental risks to human health, with one in nine deaths linked to poor indoor or outdoor air quality. The World Health Organization (WHO) estimates that 92% of the world’s population lives in locations where local air pollution exceeds WHO limits. Energy efficiency can reduce both indoor and outdoor concentrations of air pollutants. In doing so, energy efficiency drives a range of economic, environmental and health benefits associated with local air quality.

The energy system contributes vitally to economic and social progress around the world, but the associated emissions and negative side effects are costly. Scaling up the Machine Learning models to a global scale helps in proficient saving of fuel and reduces air pollution.

The remainder of the paper is organized as follows: Section 1 consists of Introduction, Section 2 consists of Literature Survey, Section 3 consists of Methodology, Section 4 consists of Results and Analysis, Section 5 consists of Conclusion and Future Scope.

2. Literature Review

The summary of the literature review can be seen in Table 1. Several approaches have been performed on this popular dataset, but the accuracy obtained by all the approaches is less than 10% RMSE.

Sr.no	Author	Year	Findings
1.	K. Hu, J. Wu, and M. Liu	2018	Modelling of EVs energy consumption from perspective of field test data using Random Forest and developing driving style questionnaires using CNN.
2.	Z. Xu, T. Wei, S. Easa, X. Zhao, and X. Qu,	2018	Modeling relationship between truck fuel consumption and driving behavior using data from internet of vehicles by integrating mobile app to the Deep Learning Module.
3.	H. Wang	2017	Energy consumption in transport: an assessment of changing trend, influencing factors and consumption forecast.
4.	D. Yang, M. Li, and X. Ban	2016	Real-time on-board monitoring method of gasoline vehicle fuel consumption based on OBD system.
5.	S. Wickramanayake and H. M. N. D. Bandara	2016	Fuel consumption prediction of fleet of vehicles using machine learning and deep learning.
6.	X.-h. Zhao, Y. Yao, Y.-p. Wu, C. Chen, and J. Rong,	2016	Prediction model of driving energy consumption based on PCA and BP network.

7.	W. J. Zhang, S. X. Yu, Y. F. Peng, Z. J. Cheng, and C. Wang	2015	Driving habits analysis on vehicle data using error back-propagation and neural network algorithm
8.	Z. Ramedani, M. Omid, A. Keyhani, S. Shamshirband, and B. Khoshnevisan,	2014	Potential of radial basis function based support vector regression for global fuel consumption.
9.	H.-l. Feng,	2013	Study on prediction model of fuel consumption index in Chongqing city based on SVR model.
10.	G. Guido, A. Vitale, V. Astarita, F. Saccomanno, V. P. Giofré, and V. Gallelli	2012	Estimation of safety performance measures from smartphone sensors and merging with neural network.
11.	D. A. Johnson and M. M. Trivedi,	2011	Driving style recognition using a smartphone as a sensor platform and neural network and back propagation network.
12.	J. N. Barkenbus,	2010	Eco-driving: an overlooked climate change initiative – fuel consumption analysis using Random Forest Modeling.
13.	T. Hiraoka, Y. Terakado, S. Matsumoto, and S. Yamabe,	2009	Quantitative evaluation of eco-driving on fuel consumption based on driving simulator experiments.
14.	K. Ahn and H. Rakha,	2008	The effects of route choice decisions on vehicle energy consumption and emissions,

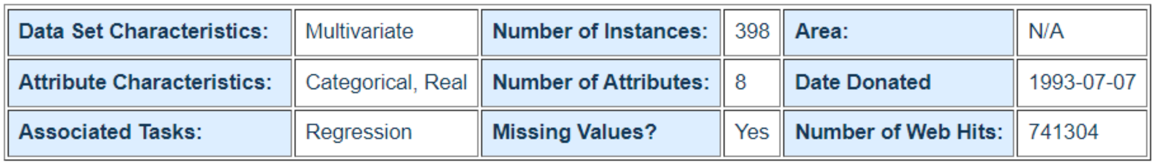
15.	Dan Pelleg.	2004	Scalable and Practical Probability Density Estimators for Fuel consumption prediction.
16.	Christopher R. Palmer and Christos Faloutsos	2003	Electricity Based External Similarity of Categorical Attributes.
17.	Thomas Melluish and Craig Saunders and Ilia Nouretdinov and Volodya Vovk and Carol S. Saunders and I. Nouretdinov V.	2001	The typicalness framework: a comparison with the Bayesian approach.
18.	Dan Pelleg and Andrew W. Moore	2001	Mixtures of Rectangles: Interpretable Soft Clustering.
19.	Zhi-Hua Zhou and Shifu Chen and Zhaoqian Chen.	2000	A Statistics Based Approach for Extracting Priority Rules from Trained Neural Networks.
20.	Mauro Birattari and Gianluca Bontempi and Hugues Bersini.	1998	Lazy Learning Meets the Recursive Least Squares Algorithm resulting in maximum mean square error.

Table 1

3.Methodology

2.1 Description of Dataset

The dataset used for this research purpose was the UCI Machine Learning Repository: Auto MPG Dataset. This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University. The dataset was used in the 1983 American Statistical Association Exposition.



Data Set Characteristics:	Multivariate	**Number of Instances:**	398	**Area:**	N/A
Attribute Characteristics:	Categorical, Real	**Number of Attributes:**	8	**Date Donated**	1993-07-07
Associated Tasks:	Regression	**Missing Values?**	Yes	**Number of Web Hits:**	741304

Schema 1

This dataset is a slightly modified version of the dataset provided in the StatLib library. In line with the use by Ross Quinlan (1993) in predicting the attribute "mpg", 8 of the original instances were removed because they had unknown values for the "mpg" attribute. The data concerns city-cycle fuel consumption in miles per gallon, to be predicted in terms of 3 multivalued discrete and 5 continuous attributes. (Quinlan, 1993)

Attribute Information:

1. mpg: continuous
2. cylinders: multi-valued discrete
3. displacement: continuous
4. horsepower: continuous
5. weight: continuous
6. acceleration: continuous

7. model year: multi-valued discrete

8. origin: multi-valued discrete

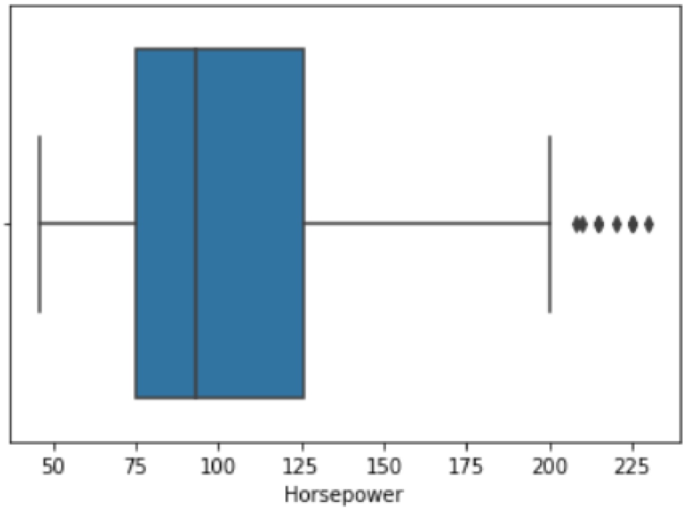
9. car name: string (unique for each instance

2.2 Preprocessing of Dataset

Framing the problem based on the dataset description and initial exploration i.e the data contains MPG variable which is continuous data and tells us about the efficiency of fuel consumption of a vehicle. The aim here is to predict the MPG value for a vehicle given other attributes of that vehicle.

2.2.1 Categorical Distribution

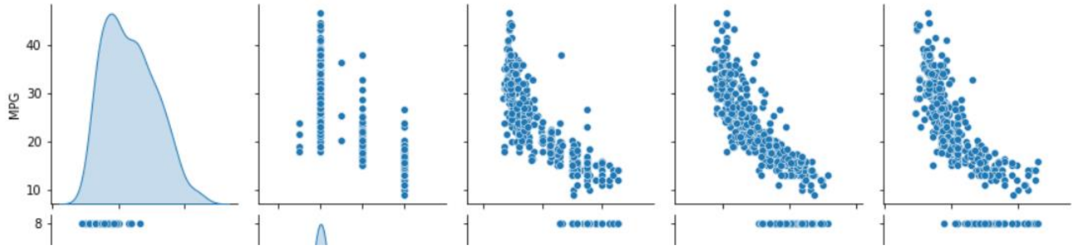
The dataset has null values. But many outliers needed to be handled properly, and also the dataset is to be categorically properly distributed. Since there are a few outliers, using the median of the column to impute the missing values by pandas median() method is the way to move forward.



Schema 2

2.2.2 Plotting for Correlation

Various plotting techniques were used for checking the distribution of data. All these preprocessing techniques play an important role when passing the data for prediction purposes. The two categorical columns are Cylinders and Origin, which only have a few categories of values. Looking at the distribution of the values among these categories tells how the data is distributed. Plotting for Correlation - The pair plot gives a brief overview of how each variable behaves with respect to every other variable.



Schema 3

For example, the MPG column (target variable) is negatively correlated with the displacement, weight, and horsepower features.

2.2.3 Hot Encoding - Checking of Origin Column

The Origin column about the origin of the vehicle has discrete values that look like the code of a country. To add some complication and make it more explicit, these numbers are converted to strings.

2.2.4 Adding Attributes using BaseEstimator and Transformer

Testing for new variables — Analyze the correlation of each variable with the target variable. Found acceleration_on_power and acceleration_on_cylinder as two new variables which turned out to be more positively correlated than the original variables.

2.2.5 Pipeline Creation

Setting up Data Transformation Pipeline for numerical and categorical attributes

As one need to automate as much as possible, Sklearn offers a great number of classes and methods to develop such automated pipelines of data transformations.

The major transformations are to be performed on numerical columns. The cascaded of set of transformations are:

1.Imputing Missing Values — using the SimpleImputer.
2.Custom Attribute Addition— using the custom attribute class.
3.Standard Scaling of each Attribute — scaling the values before feeding them to the ML model, using the standardScaler class.

2.3 Machine Learning Models

2.3.1 Machine Learning Models

The multivariate dataset consists of 3 multivalued discrete and 5 continuous attributes in addition to 2 attributes which are derived from the dataset. Regression is applied on the dataset where all the models could predict fuel consumption accurately.

Various Machine Learning models have an absolute relative error less than 10%. The random forest model is proved to have the highest accuracy and runs faster, making it suitable for wide application. This method lays a foundation for monitoring database improvement and fine management of urban – rural transportation fuel consumption.

Since Regression is applied, following models are trained:

1.Linear Regression

2.Decision Tree Regressor

3.Random Forest Regressor

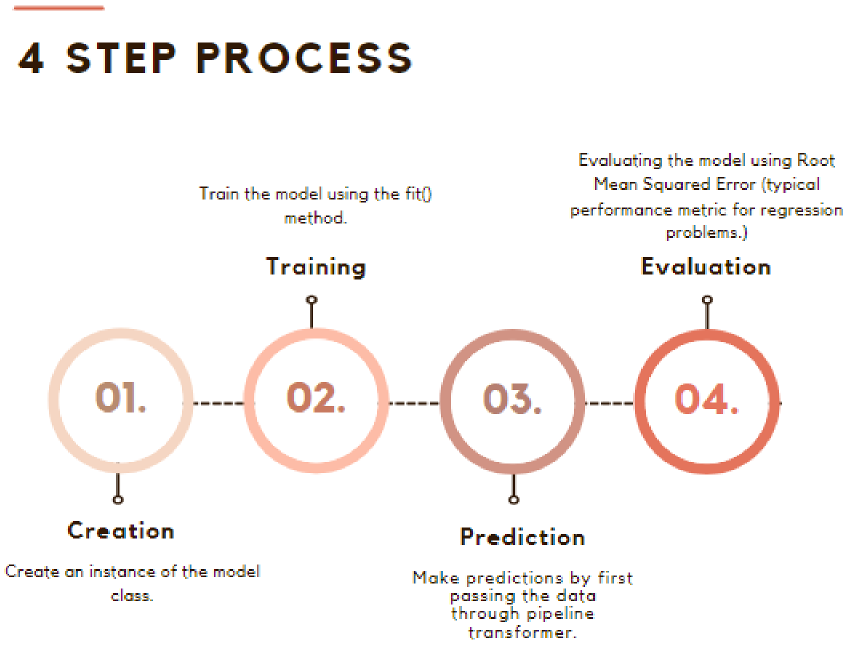
4.SVM Regressor

Now, on performing the same for Decision Tree, a 0.0 RMSE value is achieved which is not possible – as there is no “perfect” Machine Learning Model (we’ve not reached that point yet).

2.3.2 Overfitting

Problem: Testing the model on the same data one trained on, is a problem. Now, one can’t use the test data yet until the best model is finalized and is ready to go into production.

Solution: Cross-Validation



Create an instance of the model class.

Train the model using the fit() method.

Make predictions by first passing the data through pipeline transformer.

Evaluating the model using Root Mean Squared Error (typical performance metric for regression problems.)

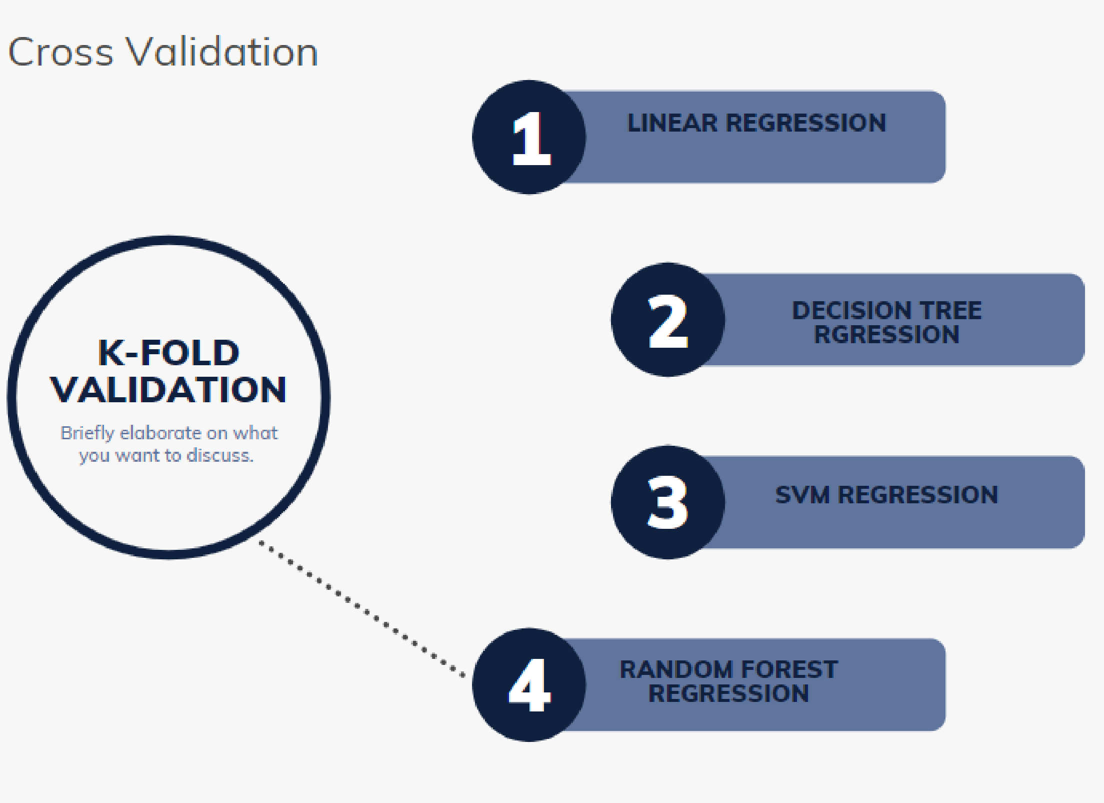
Schema 4

2.3.4 Cross Validation

Scikit-Learn’s K-fold cross-validation feature randomly splits the training set into K distinct subsets called folds. Then it trains and evaluates the model K times, picking a different fold for evaluation every time and training on the other K-1 folds.

The result is an array containing the K evaluation scores.

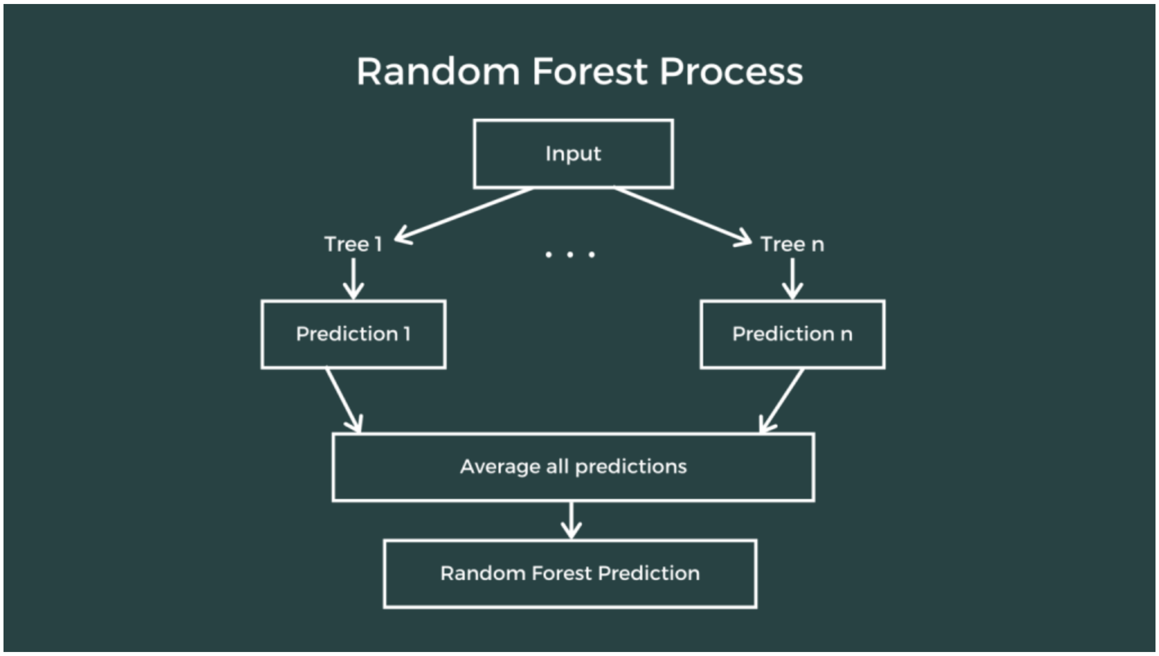
All the Machine Learning models have an absolute relative error less than 10%. The random forest model is proved to have the highest accuracy and runs faster, making it suitable for wide application.



Schema 5

2.3.5 Random Forest Regression

Every individual decision tree has high variance, but when combining all of them together in parallel then the resultant variance is low, as each individual decision tree gets perfectly trained on the sample data and hence output doesn't depend on one decision tree but multiple decision trees. In the case of a regression problem, the final output is the mean of all the outputs. This is called Aggregation. A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as bagging. The basic idea is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.



Schema 6

2.3.6 HyperParameter Tuning

After testing all the models, RandomForestRegressor has performed the best but it still needs to be fine-tuned. A model is like a radio station with a lot of knobs to handle and tune. One can either tune all these knobs manually or provide a range of values/combinations that one need to test. GridSearchCV is used to find out the best combination of hyperparameters for the RandomForest model.

2.3.7 Feature Importance

acc_on_power, which is the derived feature, has turned out to be the most important feature. One needs to keep iterating a few times before finalizing the best configuration. The model is ready with the best configuration.

4.Analysis of Results

By applying different machine learning algorithms and then using K-fold validation to see which model comes out to see which model has the least when it is applied to the data. Four models were trained. Linear Regression, Decision Tree Regression and SVM

Regression were found to have higher Root Mean Square Error than Random Forest Regression. Random Forest Regression was tested on the test data and showed promising results. The model was trained and saved as a pickle file for future deployment.

5.Conclusion and Future Scope

In this paper, four machine learning methods are proposed in which K-fold analysis was done and promising results were achieved. The conclusion found is that machine learning algorithms performed better in the analysis was Random Forest Regression Many researchers have previously suggested one should use ML and Deep Learning even for small dataset but the use of training and testing the Random Forest Regression alone proved effective results, which is proved in the paper. For the 11 features which were in the dataset, acceleration on power proved to be the estimator among the features and performed better in the Random Forest Regression ML approach when data is applied after preprocessing.

It was also found out that the dataset should be pre-processed otherwise, the feature estimator could have been missed. The Feature attribute was attained by attribute addition by using BaseEstimator and TransformerMixin. The training model gets overfitted sometimes and the accuracy achieved is not sufficient when a model is evaluated for real-world data problems which can vary drastically to the dataset on which the model was trained. Testing the model on the same data one trained on, is a problem. Now, one can't use the test data yet until best model is finalized and is ready to go into production. The K-fold Cross Validation. Random Forest regression came out to be the best and suited model.

For Future use cases and real-world datasets, use of Machine Learning Models and Deep learning i.e creation of a neural network with backtracking might produce promising results. The future of Fuel-Efficiency on real world big data sets looks promising since optimal results were shown by Forest Regression ML model on small datasets. An Upscaling with a similar approach, the future of fuel-efficiency prediction on real-world data looks bright.

References

1.Dan Pelleg. Scalable and Practical Probability Density Estimators for Scientific Anomaly Detection. School of Computer Science Carnegie Mellon University. 2004.

2.Qingping Tao Ph.D. Making Efficient Learning Algorithms with Exponentially many Features. Qingping Tao A DISSERTATION Faculty of The Graduate College University of Nebraska In Partial Fulfillment of Requirements. 2004.

3.Christopher R. Palmer and Christos Faloutsos. Electricity Based External Similarity of Categorical Attributes. PAKDD. 2003.

4.Dan Pelleg and Andrew W. Moore. Mixtures of Rectangles: Interpretable Soft Clustering. ICML.

5.Jinyan Li and Kotagiri Ramamohanarao and Guozhu Dong. Combining the Strength of Pattern Frequency and Distance for Classification. PAKDD. 2001.

6.Thomas Melluish and Craig Saunders and Ilia Nouretdinov and Volodya Vovk and Carol S. Saunders and I. Nouretdinov V. The typicalness framework: a comparison with the Bayesian approach. Department of Computer Science. 2001.

7.Wai Lam and Kin Keung and Charles X. Ling. PR 1527. Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong. 2001.

8.Zhi-Hua Zhou and Shifu Chen and Zhaoqian Chen. A Statistics Based Approach for Extracting Priority Rules from Trained Neural Networks. IJCNN . 2000.

9.Mauro Birattari and Gianluca Bontempi and Hugues Bersini. Lazy Learning Meets the Recursive Least Squares Algorithm. NIPS. 1998.

10.D. Greig and Hava T. Siegelmann and Michael Zibulevsky. A New Class of Sigmoid Activation Functions That Don't Saturate. 1997.

11.Johannes Furnkranz. Pairwise Classification as an Ensemble Technique. Austrian Research Institute for Artificial Intelligence.