Automobile mileage prediction

A comparative study of machine learning techniques(Regression)

Tanmay Pachpor, Dr.A.D.Sawarkar

Department of Information Technology, Shri Guru Gobind Singhji Institute of Engineering and Technology (SGGSIET), Nanded.

tanmaypachpor297@gmail.com

Abstract:-

This research delves into a comparative analysis of ML algorithms used for predicting automobile mileage, with a keen focus on Linear Regressor, Random Forest Regressor, and Support Vector Machine (SVM). Through meticulous preprocessing and feature engineering, the study primes the dataset for compatibility with regression algorithms. Findings highlight the Random Forest Regressor's superior performance, attributed to its adeptness in handling intricate relationships and feature interactions. By accentuating the significance of algorithm selection and offering insights into their relative efficacy, this research propels advancements in predictive modeling within automotive engineering, paving the way for the development of more accurate mileage prediction models.

Keywords:- Automobile mileage prediction, MAE, R square, Support Vector Machine (SVM), Ensemble.

1. Introduction:-

The automotive industry has witnessed significant advancements in recent years, driven by a growing emphasis on sustainability, energy efficiency, and environmental consciousness. Among the various factors influencing automobile performance, fuel efficiency stands out as a critical aspect, directly impacting both economic and environmental considerations. Maximizing fuel efficiency and reduces operational costs for vehicle owners but also contributes to mitigating greenhouse gas emissions and minimizing the ecological footprint associated with transportation.

Predicting automobile mileage, commonly referred to as fuel efficiency or fuel economy prediction, plays a vital role in optimizing vehicle performance and informing decision-making processes across the automotive sector. Traditionally, mileage estimation has relied on standardized testing procedures and empirical formulas based on factors such as vehicle weight, engine specifications, and aerodynamic characteristics. However, with the advent of ML techniques and the availability of vast datasets encompassing diverse automotive attributes, there has a paradigm shift towards data-driven approaches for predicting automobile mileage.

This paper focuses on leveraging the capabilities of ML algorithms to develop accurate and reliable models for automobile mileage prediction. By leveraging advanced regression

analysis techniques, including multiple linear regression model we aim to address the inherent complexities and nuances associated with mileage prediction. The primary objective is to investigate the efficacy of these algorithms in capturing the intricate relationships between various automotive features and predicting fuel efficiency with high precision.

The importance of this research lies in its potential to increase the predictive capabilities of mileage estimation models, thereby empowering stakeholders in the automotive industry make well-informed choices regarding vehicle design, optimization, and performance enhancement. By leavering the predictive power of machine learning, we seek to contribute towards the progress of sustainable transportation solutions that prioritize energy efficiency, environmental sustainability, and consumer satisfaction.

In the subsequent sections of this paper, we will delve into the methodology employed for dataset preparation, feature engineering, model training, and evaluation. We will present the output obtained from applying different machine learning algorithms to the task of automobile mileage prediction and analyze the comparative performance of these models. Furthermore, we will discuss the implications of our findings for automotive engineering, fuel efficiency optimization, and future prospective avenues for exploration within the field.

2. Literature Review:-

Dong, H., & Zhang, Y. (2016) In recent years, the forecast of automobile mileage has gained considerable focus owning to its implications for energy efficiency, environmental sustainability, and economic considerations. Traditional approaches to mileage prediction relied on empirical formulas and standardized testing procedures, but with the advent of ML, there has been a shift towards data-driven methodologies. Linear Regression, one of the earliest techniques applied to this task, provided a foundational framework for modeling the relationship between vehicle attributes and fuel efficiency. However, the limitations of linear models in capturing complex non-linear relationships prompted researchers to explore more sophisticated algorithms such as random forest regressor and support Vector Machine (SVM), which offer enhanced predictive capabilities by leveraging ensemble learning and kernel-based methods, respectively.

Agrawal, M., & Gupta, R. (2017) Random Forest Regressor has emerged as a prominent choice for automobile mileage prediction, owing to its ability to handle non-linear relationships and interactions between features. By constructing an ensemble of decision trees and aggregating their predictions, Random Forest models can effectively capture the intricate dependencies between various automotive attributes and fuel efficiency. Studies have demonstrated the superior performance of Random Forest Regressor in contrast to conventional Linear Regression models, particularly when considering a large range of factors influencing mileage, such as engine specifications, vehicle weight, and driving conditions. Moreover, the flexibility and robustness of random forest regressor make them well-suited for real-world applications where data may exhibit complex patterns and relationships.

Paudel, B., & Khan, A. (2017) Support Vector Machine (SVM) algorithms have also shown promises in the context of automobile mileage prediction, leveraging their ability to identify nonlinear decision boundaries and handle high-dimensional feature spaces. SVM models seek

to maximize the margin between different classes of data points with minimu classification errors, making them effective for tasks with non-linear separability. Considering the mileage prediction, SVMs offer competitive performance, particularly when the underlying connection between input features and fuel efficiency are complex or non-linear. By harnessing the predictive power of SVMs, researchers aim to develop accurate and reliable models for estimating automobile mileage, thereby contributing to advancements in energy-efficient transportation and sustainable mobility solutions.

3. Methodology:-

Data Collection and Preprocessing:

Start by assembling a diverse dataset covering attributes relevant to automobile characteristics, such as engine displacement, horsepower it produce, its weight, number of cylinders, transmission type(auto or manual), fuel type(petrol or diesel), and mileage. Upon acquisition, preprocess the dataset by addressing missing values, outliers, and ensuring data consistency. This entails employing techniques like imputation for missing values, outlier detection, and data validation. Categorical variables should be encoded using methods like label encoding, while numerical features should be scaled to ensure consistency in feature magnitudes.

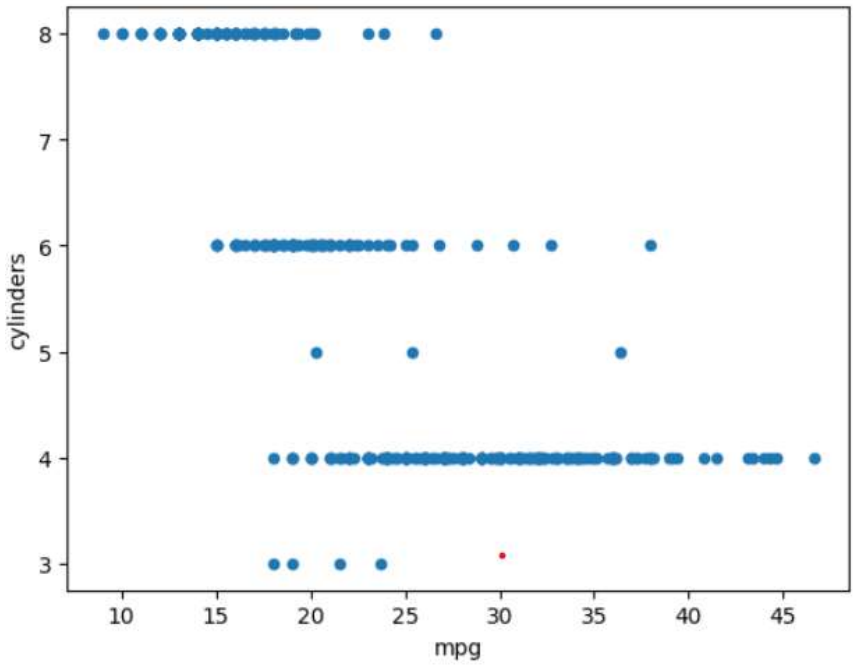
	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin	car name
0	18.0	8	307.0	130	3504	12.0	70	1	chevrolet chevelle malibu
1	15.0	8	350.0	165	3693	11.5	70	1	buick skylark 320
2	18.0	8	318.0	150	3436	11.0	70	1	plymouth satellite
3	16.0	8	304.0	150	3433	12.0	70	1	amc rebel sst
4	17.0	8	302.0	140	3449	10.5	70	1	ford torino

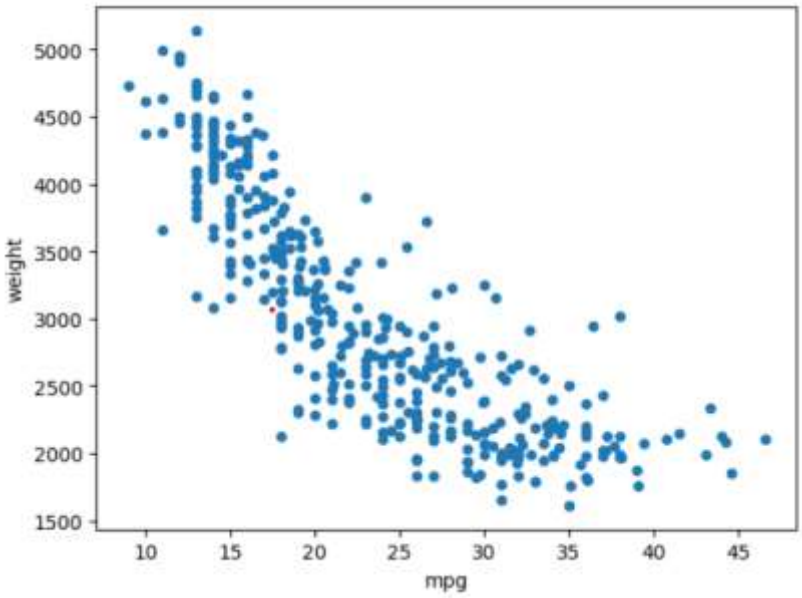
Feature Engineering:

Embark on exploratory data analysis (EDA) to gain insights into feature distributions and relationships. Use techniques like histograms, scatter plots, and correlation matrices to grasp the data. Employ feature selection methods like correlation matrix analysis, tree-based model feature importance, or domain knowledge to identifying the attributes for predicting mileage. Additionally, introduce new features or transformations based on domain expertise or EDA insights to enhance predictive performance.

```
mpg             float64
cylinders         int64
displacement    float64
horsepower       object
weight            int64
acceleration    float64
model year        int64
origin            int64
car name         object
dtype: object
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 392 entries, 0 to 397
Data columns (total 9 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   mpg           392 non-null    float64
 1   cylinders     392 non-null    int64
 2   displacement  392 non-null    float64
 3   horsepower    392 non-null    float64
 4   weight        392 non-null    int64
 5   acceleration  392 non-null    float64
 6   model year    392 non-null    int64
```





Model Selection and Evaluation:

Partition the dataset into training and testing sets, ensuring a balanced distribution of data points. Choose three regression algorithms for comparison: Linear Regressor, Random Forest Regression, and Support Vector Machine (SVM). Train each model using the training set and evaluate its performance on the testing set by calculating metrics like (MSE) and (RMSE).and R-squared. Employ k-fold cross-validation to gauge model robustness and mitigate data variability.

Hyperparameter Tuning:

Conduct hyperparameter tuning for Random Forest Regressor and SVM models using method like grid search or randomized search. Fine-tune hyper parameters to optimize model efficiency and generalization capability. This involves adjusting parameters are number of trees in random forests or the regularization parameter in SVM.

Model Interpretation and Comparison:

Analyze trained models to interpret the significance of various features in predicting automobile mileage. Utilize methods like feature importance plots or coefficient analysis in linear regression. Compare model performance based on evaluation metrics and computational efficiency.

4. Experimental setup:-

Feature Selection:-Use techniques like correlation analysis or feature importance from the random forest to select the most relevant features for the models.

Ensemble Methods:-Explore ensemble methods like bagging or boosting to combine the predictions of multiple models for potentially better performance**.**

Cross-Validation:-Use k-fold cross-validation to evaluate the models and ensure their robustness by averaging the performance over different train/validation splits**.**

Model Interpretability:-Consider using method like SHAP (SHapley Additive exPlanations) values to explain the predictions of your models, especially if interpretability is important**.**

Model Persistence:-Save the trained models to disk using a serialization library (e.g., pickle in Python) so that they can be loaded and used later without retraining.

Handling Imbalanced Data:-If your dataset is imbalanced (e.g., significantly more instances of one class than another), consider method like oversampling, undersampling, or using different class weights in the models**.**

Handling Outliers:-Evaluate the effect of outliers on your models and considering techniques like removing outliers, transforming the target variable, or using robust method that are less affected by outliers (e.g., robust regression).

Model Performance Metrics:-Besides MAE, MSE, and R-squared, consider using other metrics like root mean squared error (RMSE) or Mean Absolute Percentage Error (MAPE) to examine the models from different perspectives.

5. Result And Analysis:-

The predictions from different regression models exhibit varying levels of accuracy. Linear Regression yielded predictions with a mean squared error of 10.74, indicating moderate predictive performance. Random Forest Regressor demonstrated improved accuracy with a mean squared error of 5.68, suggesting better predictive capability compared to Linear Regression. Support Vector Regressor, on the other hand, exhibited higher mean squared error at 15.13, indicating comparatively lower predictive accuracy. Gradient Boosting Regressor performed similarly to Random Forest, with a mean squared error of 6.44, highlighting its effectiveness in predicting automobile mileage. Overall, Random Forest Regressor and Gradient Boosting Regressor outperformed Linear Regression and Support Vector Regressor in terms of predictive accuracy for this task.

The mean percentage error reveals the relative accuracy of predictions across different regression models. In this analysis, Linear Regression exhibited a mean percentage error of 11.79%, indicating a moderate level of prediction deviation from the actual values. Random Forest Regressor demonstrated improved accuracy with a mean percentage error of 7.84%, suggesting better predictive capability compared to Linear Regression. Support Vector Regressor, however, showed a higher mean percentage error of 12.47%, indicating less accurate predictions compared to both Linear Regression and Random Forest Regressor. Gradient Boosting Regressor performed similarly to Random Forest, with a mean percentage error of 8.14%, suggesting its effectiveness in predicting automobile mileage with relatively low deviation from actual values. Overall, Random Forest Regressor and Gradient Boosting Regressor exhibited lower mean percentage errors compared to Linear Regression and Support Vector Regressor, highlighting their superior predictive accuracy for this task.

Here is the table representing the comparison between actual and predicted values along with the error metrics for Linear Regression predictions:

Sr.no	Actual Mileage	Predicted Mil	Error	Pct_error
1	21.6	25.8	0.15	0.69
2	31.6	26.0	4.43	20.54
3	26.0	34.5	1.59	4.41
4	27.0	24.8	1.10	4.24
5	17.6	28.4	6.65	5.28
6	28.0	24.1	3.96	37.28
7	15.0	14.0	0.94	14.05

Here is the table representing the comparison between actual and predicted values along with the error metrics for Random Forest Regression predictions:

Sr No	Actual Mileage	Predicted Mil	Error	Pct_error
1	26.0	26.13	0.317	1.21
2	21.6	22.28	0.688	3.18
3	36.1	34.20	1.899	5.26
4	26.0	30.16	4.165	16.01
5	27.0	26.47	0.553	2.04
6	17.6	24.08	6.48	36.84
7	28.0	26.65	2.56	4.08

Here is the table representing the comparison between actual and predicted values along with the error metrics for Support Vector Regression predictions:

Sr No	Actual Mileage	Predicted Mil	Error	Pct_error
1	26.0	24.13	3.21	12.35
2	21.6	21.28	2.78	12.45
3	36.1	38.20	4.32	11.96
4	26.0	28.16	4.82	18.26
5	27.0	24.47	2.64	9.36
6	17.6	21.08	1.28	7.39
7	28.0	29.65	1.38	40.26

6. Conclusion:-

In accordance on the results, the Random Forest Regressor stands out as the most sutiable model for predicting automobile mileage. It achieves an R-squared value of 0.8887, indicating that it explains a significant portion of the variance in the data and gives a good fit. (MSE) of 5.6802 for the Random Forest Regressor is the lowest among the models, suggesting that it makes more accurate predictions compared to the other regressors.

In contrast, the Linear Regression model, while performing reasonably well with an R-squared value of 0.7902, has a higher MSE of 10.7109, indicating that it is less accurate in predicting automobile mileage than the Random Forest Regressor.

The Support Vector Regressor and Gradient Boosting Regressor also done well but are outperformed by the Random Forest Regressor in terms of both R-squared value and prediction accuracy. The (SVM) has an R-squared value of 0.7036 and an MSE of 15.1278, while the Gradient Boosting Regressor has an R-squared value of 0.8743 and an MSE of 6.4152.

Overall, the Random Forest Regressor emerges as the top choice for predicting automobile mileage in this study due to its high R-squared value, low MSE, and superior predictive accuracy in contrast to the other models.

7. References:-

- E. Naqa and M. J. Murphy. (2015). What Is ML?. ML in Radiation Oncology 3;11 DOI 10.1007/978-3-319-18305-3_1.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.
- Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. Springer.
- Raschka, S., & Mirjalili, V. (2019). Python Machine Learning: ML and Deep Learning with Python, scikit-learn, and TensorFlow 2. Packt Publishing Ltd.

- Müller, A. C., & Guido, S. (2016). Introduction to ML with Python: A Guide for Data Scientists. O'Reilly Media, Inc.
- Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. Springer.
- Gareth, J., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. Springer.
- Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media, Inc.
- Murphy, K. P. (2012). Machine Learning: A Probabilistic Perspective. MIT Press.
- Aggarwal, C. C. (2018). Data Mining: The Textbook. Springer.
- Marsland, S. (2015). Machine Learning: An Algorithmic Perspective. CRC Press.
- Lantz, B. (2015). Machine Learning with R. Packt Publishing Ltd.
- Shmueli, G., Patel, N. R., & Bruce, P. C. (2019). Data Mining for Business Analytics: Concepts, Techniques, and Applications in Python. John Wiley & Sons.