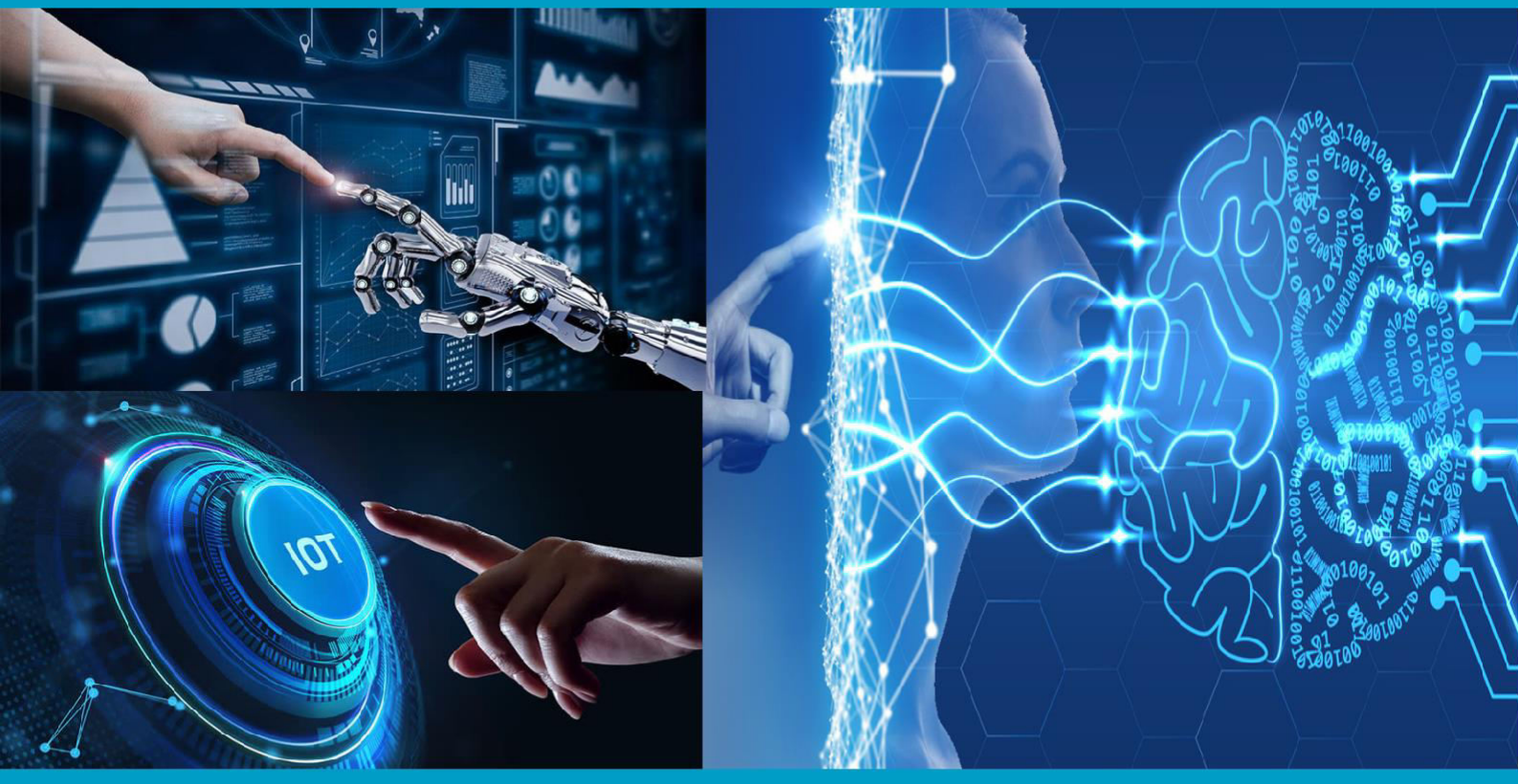
IJMRSETM

e-ISSN: 2395 - 7639



INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING, TECHNOLOGY AND MANAGEMENT

Volume 11, Issue 6, June 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.802

International Journal of Multidisciplinary Research in Science, Engineering, Technology & Management (IJMRSETM)

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 7.802 | A Monthly Double-Blind Peer Reviewed Journal |

| Volume 11, Issue 6, June 2024 |

Car Price Prediction Using Machine Learning Techniques

Rohit Dattatray Sonawale, Prajesh Prakash Makootan, Rutuja Chavan, Divakar Jha

Department of MCA, Late Bhausaheb Hiray S. S. Trust's Institute of Computer Application, Mumbai, India

ABSTRACT: In this paper, we investigate the application of machine learning techniques to predict car prices using the Quikr dataset. The study employs various machine learning algorithms, such as linear regression to estimate car prices based on features such as name, company, year, Price, kms_driven, fuel_type. Our analysis demonstrates the comparative performance of these models and identifies the most influential features affecting car prices. The findings indicate that machine learning models can provide accurate and reliable price predictions, offering valuable insights for buyers and sellers in the online car marketplace. This research underscores the potential of predictive analytics in enhancing the efficiency and transparency of the automotive market.

KEYWORDS: Supervised Machine Learning Algorithm, Linear Regression, Predictive Analytics

I. INTRODUCTION

The automotive industry in India is the largest market for both foreign and domestic automakers. A sizable market for used cars is developing along with the expansion and demand for autos. Some internet advertising websites, such as Olx and Quickr, are manipulating and controlling the used automobile market. However, buyers of secondhand cars are easily tricked and influenced into paying a price that is too high for the vehicle. I would like to suggest utilizing supervised learning machine learning techniques and algorithms to forecast used car values depending on certain parameters in order to solve this challenge with the use of artificial intelligence and machine learning. And I'd like to look at and compare the accuracy that various algorithms yield when testing and forecasting using data from previously owned cars. In India, 26,353,293 cars were produced overall in 2019–20; nevertheless, 22,652,108 cars were produced there in 2020–21. For this reason, I set out to solve this issue and identify a prediction approach that would provide used car pricing in an accurate manner. This automobile price prediction experiment uses data from a variety of websites, including open-source data websites, web scraping, and Kaggle, which offers free data. Given that it necessitates a discernible amount of work and expertise from the subject matter expert, automobile price prediction research has garnered a lot of interest. For a dependable and accurate prediction, a large number of unique attributes are looked at. The automobile is worth the asking price. With machine learning, the right and fair price for a particular used automobile may be predicted using historical data from different vendors and purchasers. The process involves training the model with a dataset of used cars that contains various features and parameters, including the name, company, year, price, kms_driven, fuel_type, etc. These features allow for the prediction of the car's price. Additionally, we might mention if there has been any damage, whether from flooding or unintentional harm, to these elements in an automobile, which can also be taken into account for estimating the precise and accurate pricing of the vehicle.

II. LITERATURE REVIEW

In the paper, the focus is on understanding the landscape of machine learning applications in car price prediction, specifically using the Quikr dataset and linear regression. Car price prediction is critical in the automotive industry, influencing decisions for buyers, sellers, and dealers alike. Machine learning has emerged as a powerful tool in this domain, allowing for the analysis of historical data to forecast prices accurately. Previous studies have demonstrated the effectiveness of regression models, particularly linear regression, in capturing relationships between car attributes (such as mileage, model year, brand, and condition) and market prices. Researchers have often utilized various datasets, each offering unique challenges and insights into predicting car prices.

Linear regression stands out due to its simplicity and interpretability, making it accessible for modeling the price dynamics of used cars based on quantitative features. However, the approach also faces challenges such as assumptions of linearity, the need for careful feature selection, and addressing potential outliers and data quality issues. Recent advancements in machine learning techniques, including ensemble methods and feature engineering, have

contributed to enhancing the predictive accuracy of linear regression models in this context. The Quikr dataset, known for its comprehensive listings and diverse car attributes, provides a specific and valuable resource for exploring these methodologies and further advancing the field of car price prediction. This literature review sets the stage for investigating how linear regression can be effectively applied to the Quikr dataset, offering insights into both the methodology and practical implications for stakeholders in the automotive market.

III. PROBLEM STATEMENT

Predicting car prices accurately is crucial in the automotive industry for enabling informed decisions by buyers, sellers, and dealers. However, the variability in car attributes and market dynamics makes this task challenging. Existing approaches often rely on manual assessment or simplistic pricing models, which may not capture the full complexity of factors influencing car prices. This study aims to address these challenges by developing and validating a machine learning model based on linear regression techniques. Specifically, we will leverage the Quikr dataset, known for its extensive listings and diverse set of car features, to train and evaluate the model. The primary objective is to build a robust predictive tool that can effectively estimate car prices based on key variables such as mileage, model year, brand, and condition. By exploring the applicability of linear regression in this context, we aim to provide insights into how such models can enhance price transparency and efficiency in the used car market. This research seeks to contribute to the field by advancing methodologies for car price prediction and offering practical implications for stakeholders seeking reliable valuation methods in automotive transactions.

IV. OBJECTIVE SCOPE

This research aims to achieve several objectives in the realm of machine learning-based car price prediction using the Quikr dataset and linear regression. Firstly, the primary objective is to develop a robust predictive model that accurately estimates the prices of used cars based on various features available in the Quikr dataset, such as mileage, model year, brand, and condition. The model will be trained using historical data from Quikr, which includes a diverse range of vehicles and comprehensive attribute listings. Secondly, the study seeks to validate the performance of the linear regression model through rigorous evaluation metrics, assessing its predictive accuracy and reliability in real-world scenarios. Additionally, the research will explore the impact of different feature combinations and preprocessing techniques on model performance to enhance its effectiveness in predicting car prices. By achieving these objectives, this study aims to contribute to the advancement of machine learning applications in the automotive industry, providing insights that can benefit stakeholders such as buyers, sellers, and market analysts in making informed decisions based on more accurate price estimations. The scope of this research is focused on leveraging linear regression as a fundamental predictive tool and utilizing the Quikr dataset to explore and validate its applicability in the context of car price prediction, thereby offering practical implications for enhancing transparency and efficiency in the used car market.

V. RESEARCH METHODOLOGY

a) Analysis of Quikr dataset:

In the given dataset of Quikr we have 892 rows and 6 columns respectively. In this dataset there are important features as name, company, year, Price, kms_driven, fuel_type.

There are 3 Categorical Features

- Name
- Company
- Fuel_type

There are 3 Numerical Features

- Year
- Price
- Kms_given

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 7.802 | A Monthly Double-Blind Peer Reviewed Journal |

| Volume 11, Issue 6, June 2024 |

Details of the dataset

Details of Dataset	Value
Number of Rows	892
Number of Columns	6
Number of Categorical Features	3
Number of Numerical Features	3
Are there any missing values	yes

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 892 entries, 0 to 891
Data columns (total 6 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   name        892 non-null    object
 1   company     892 non-null    object
 2   year        892 non-null    object
 3   Price       892 non-null    object
 4   kms_driven  840 non-null    object
 5   fuel_type   837 non-null    object
dtypes: object(6)
memory usage: 41.9+ KB
```

b) Pre-processing:

In the data preprocessing section of your research paper on machine learning car price prediction using the Quikr dataset, focus on the steps undertaken to prepare the data for building an effective linear regression model. The dataset comprises features such as 'name', 'company', 'kms driven', 'year', 'price', and 'fuel type'. Initially, address how missing values were handled to ensure the dataset's completeness; for instance, missing values in 'kms driven' and 'year' were imputed using the median value of these columns, while missing categorical data like 'company' and 'fuel type' were filled with the mode. Next, discuss the transformation of categorical variables into numerical form, essential for linear regression models. One-hot encoding was applied to 'company' and 'fuel type' to convert these categorical features into binary vectors. Additionally, the 'name' feature, which often provides limited predictive power, was dropped to streamline the dataset. Normalization or scaling of numerical features is another crucial step. 'Kms driven' and 'year' were standardized using Min-Max scaling to bring all feature values into a comparable range, thereby improving the model's performance. Outlier detection and removal were also conducted to eliminate any anomalous data points that could negatively impact the regression model's accuracy. Finally, discuss any feature engineering efforts undertaken to enhance the dataset, such as creating new variables that capture interactions between existing features. By detailing these preprocessing steps, you ensure that the dataset is clean, well-structured, and primed for accurate car price prediction using linear regression.

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 7.802 | A Monthly Double-Blind Peer Reviewed Journal |

| Volume 11, Issue 6, June 2024 |

c) Cleaned data:

	name	company	year	Price	kms_driven	fuel_type
0	Hyundai Santro Xing	Hyundai	2007	80000	45000	Petrol
1	Mahindra Jeep CL550	Mahindra	2006	425000	40	Diesel
2	Hyundai Grand i10	Hyundai	2014	325000	28000	Petrol
3	Ford EcoSport Titanium	Ford	2014	575000	36000	Diesel
4	Ford Figo	Ford	2012	175000	41000	Diesel
...	...	...	...	...	...	...
811	Maruti Suzuki Ritz	Maruti	2011	270000	50000	Petrol
812	Tata Indica V2	Tata	2009	110000	30000	Diesel
813	Toyota Corolla Altis	Toyota	2009	300000	132000	Petrol
814	Tata Zest XM	Tata	2018	260000	27000	Diesel
815	Mahindra Quanto C8	Mahindra	2013	390000	40000	Diesel

816 rows × 6 columns

d) Exploratory Data Analysis (EDA):

With the aid of exploratory data analysis, a statistical methodology is used to analyze data sets in order to highlight their key features through the use of statistical graphics and other data visualization techniques. Whether or whether a statistical model is employed, the main goal of EDA is to discover additional insights from the data that go beyond the scope of formal modelling or hypothesis testing.

e) One-Hot Encoding:

For categorical variables where no ordinal relationship exists, the integer encoding might not be enough, at best, or misleading to the model at the worst. during this case, a one-hot encoding may be applied to the ordinal representation. this can be where the integer encoded variable is removed and one new binary variable is added for every unique integer value within the variable. Each bit represents a possible category. If the variable cannot belong to multiple categories directly, then just one bit within the group is “on.” this can be called one-hot encoding.

f) Implementation of ML Model:

Linear Regression

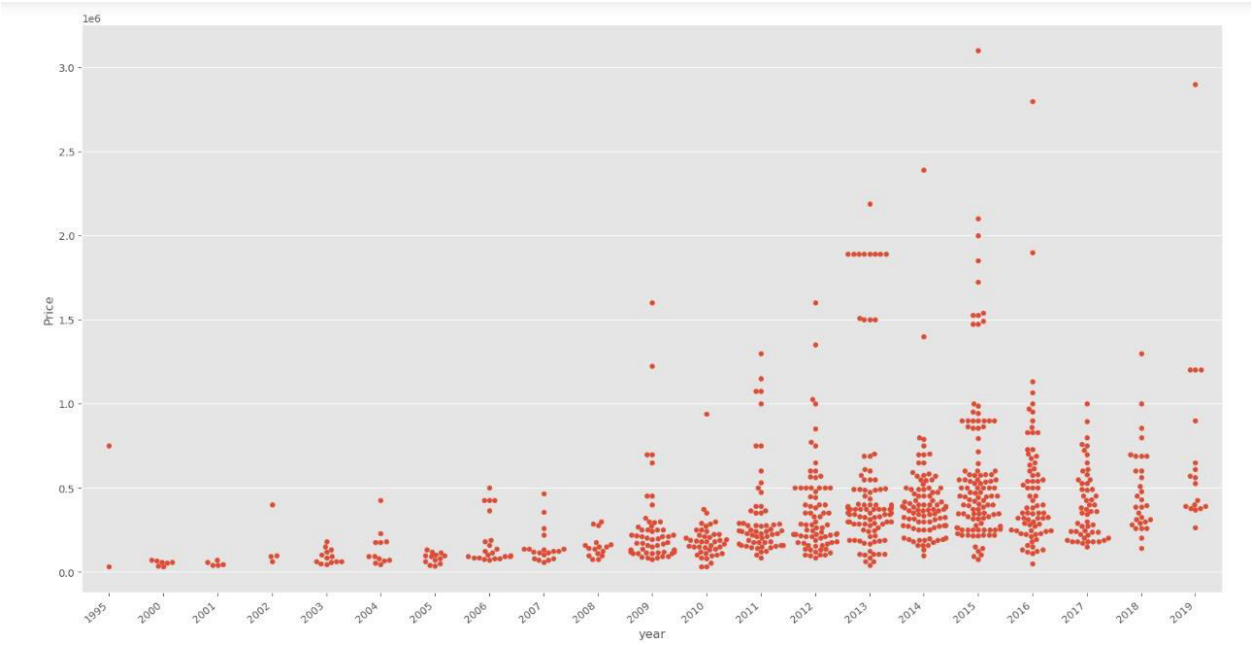
Linear regression is a fundamental machine learning algorithm used for predicting a continuous target variable, such as car prices, based on one or more predictor variables (features). The model aims to establish a linear relationship between the dependent variable (price) and the independent variables (e.g., company, kilometers driven, year, and fuel type). Mathematically, linear regression represents this relationship with an equation of the form $y=\beta0+\beta1x1+\beta2x2+\cdots+\beta nxn+\epsilon$, where y is the predicted price, β0 is the intercept, β1,β2,…,βn are the coefficients for each predictor, x1,x2,…,xn and ϵ is the error term. The model's coefficients are determined through a method called Ordinary Least Squares (OLS), which minimizes the sum of the squared differences between the observed and predicted values. Linear regression is appreciated for its simplicity, interpretability, and efficiency; however, it assumes a linear relationship between variables, is sensitive to outliers, and can be affected by multicollinearity among predictors. Despite these limitations, it remains a powerful tool for car price prediction when the relationships are approximately linear and the data is appropriately preprocessed.

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 7.802 | A Monthly Double-Blind Peer Reviewed Journal |

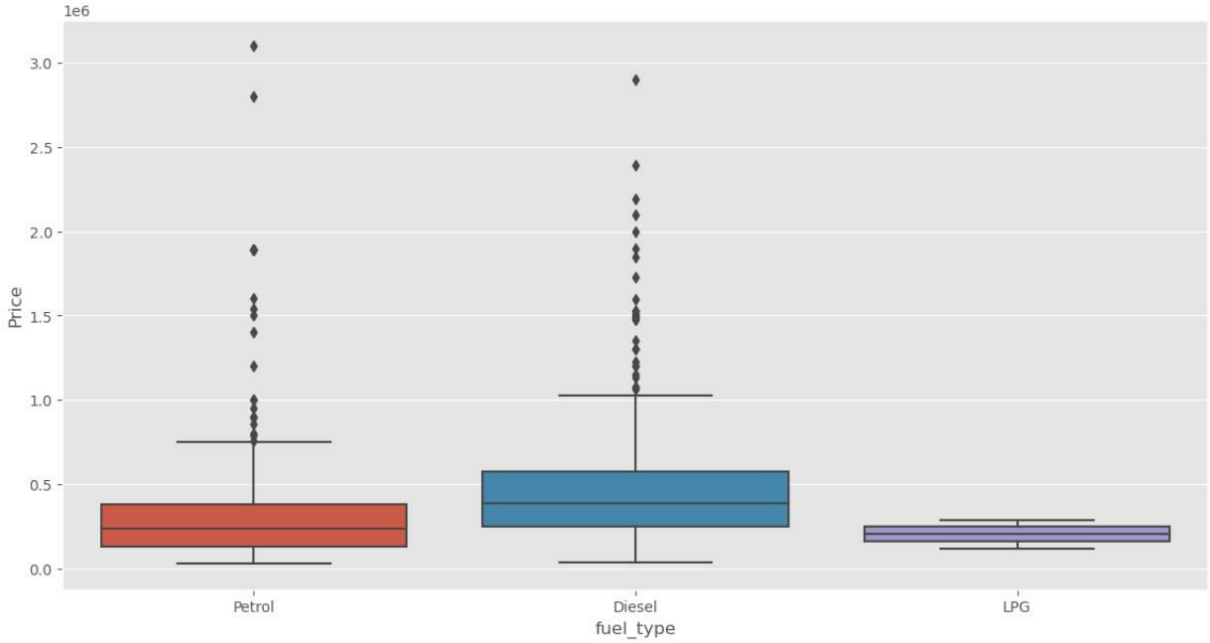
| Volume 11, Issue 6, June 2024 |

VI. ANALYSIS AND FINDINGS

Checking relationship of Year with Price



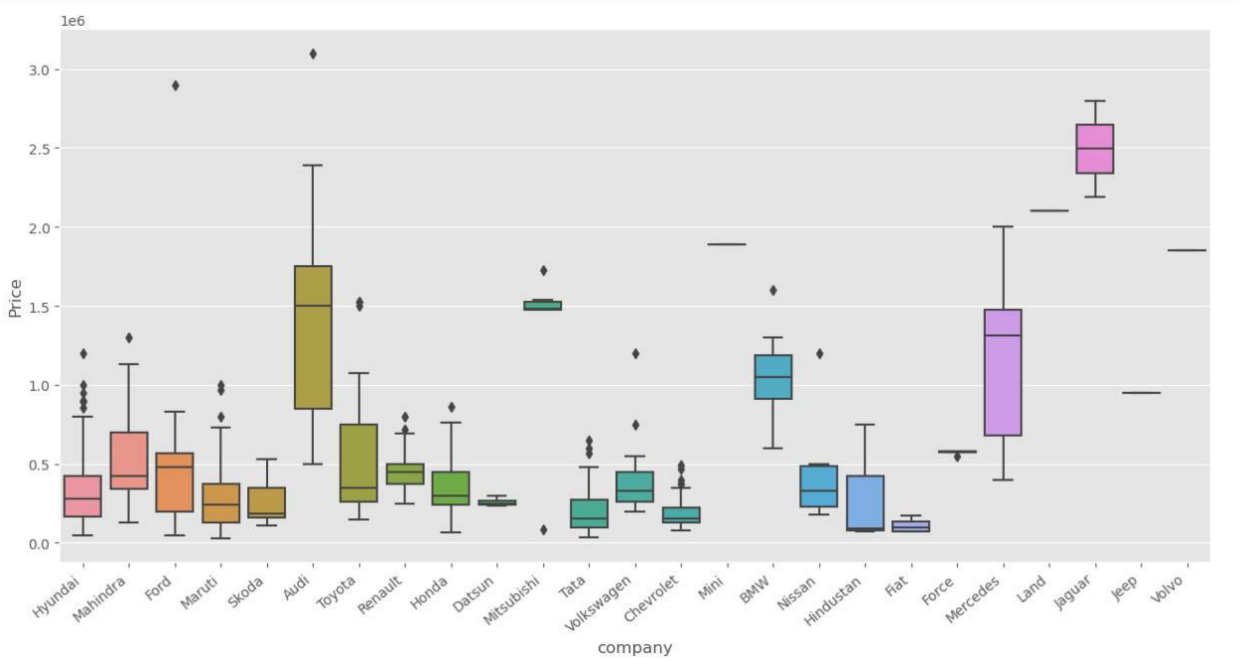
Checking relationship of Fuel Type with Price



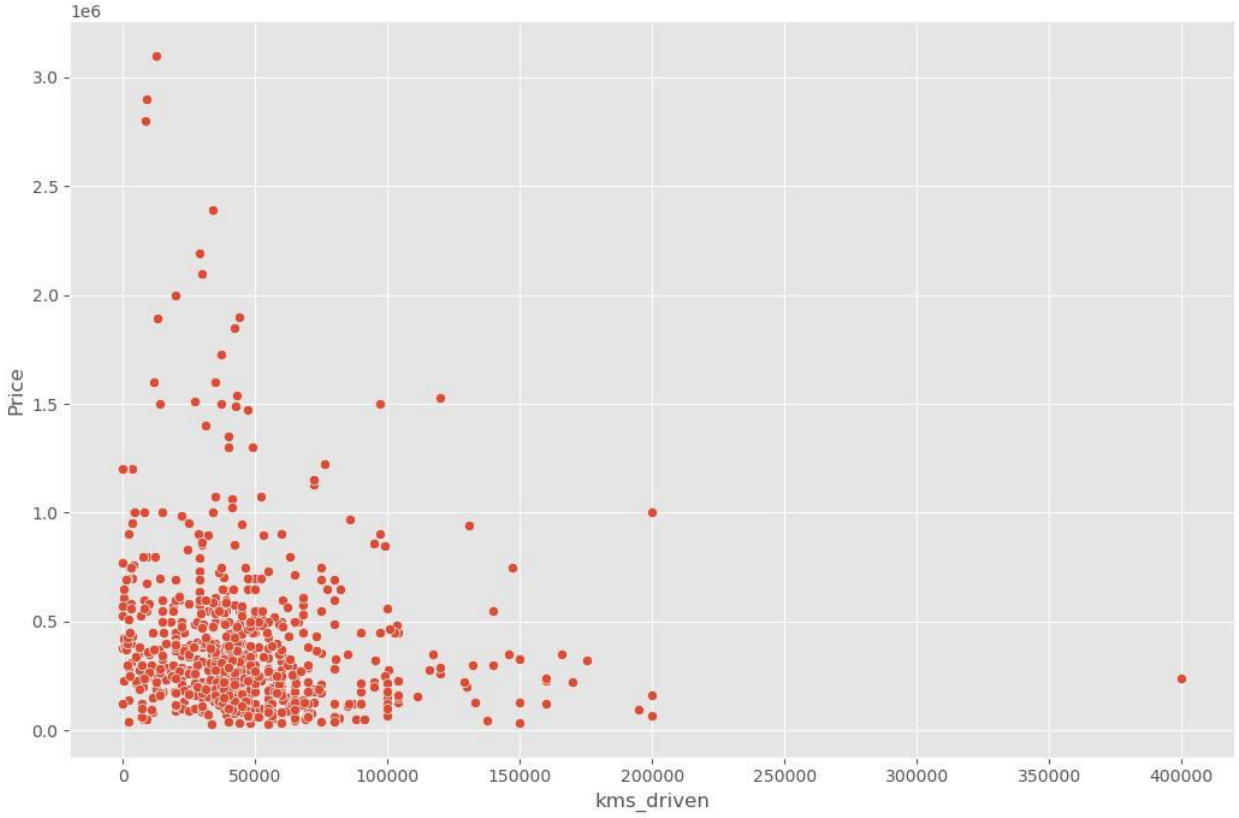
| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 7.802 | A Monthly Double-Blind Peer Reviewed Journal |

| Volume 11, Issue 6, June 2024 |

Checking relationship of Company with Price



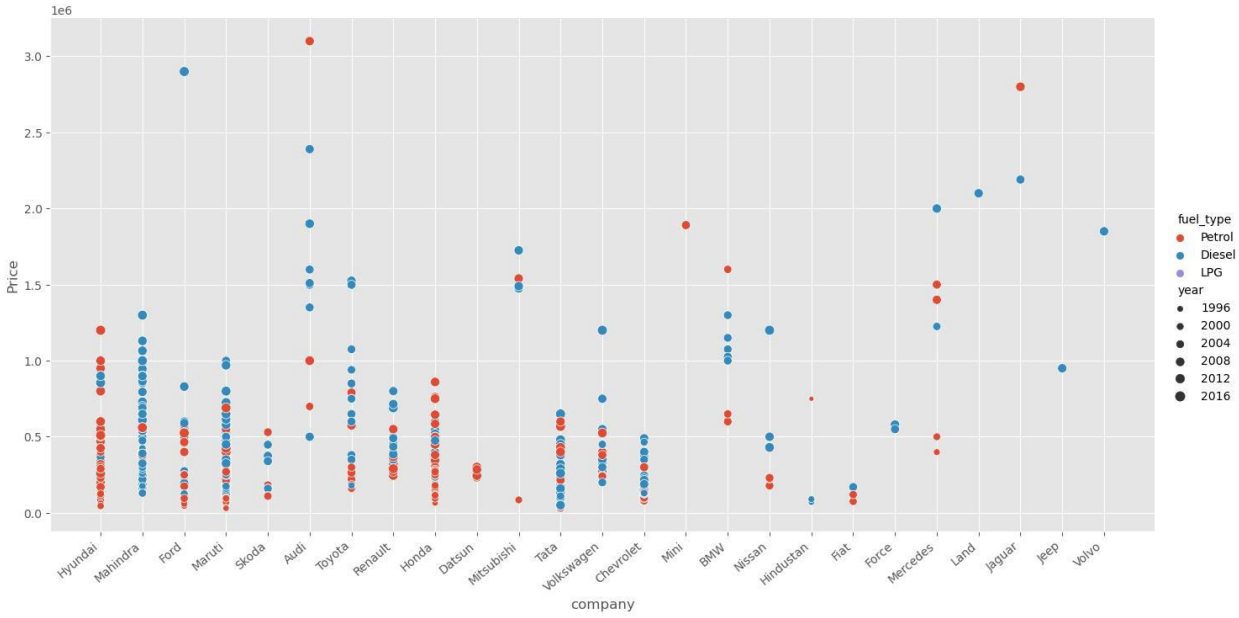
Checking relationship of Kms_driven with Price



| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 7.802 | A Monthly Double-Blind Peer Reviewed Journal |

| Volume 11, Issue 6, June 2024 |

Checking relationship of Fuel Type, Year and Company



Applying Train and Test Split

```
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.2)
```

```
from sklearn.linear_model import LinearRegression
```

```
from sklearn.preprocessing import OneHotEncoder
from sklearn.compose import make_column_transformer
from sklearn.pipeline import make_pipeline
from sklearn.metrics import r2_score
```

The Best Model

```
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.1,random_state=np.argmax(scores))
lr=LinearRegression()
pipe=make_pipeline(column_trans,lr)
pipe.fit(X_train,y_train)
y_pred=pipe.predict(X_test)
r2_score(y_test,y_pred)
```

```
0.8959285359819742
```

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 7.802 | A Monthly Double-Blind Peer Reviewed Journal |

| Volume 11, Issue 6, June 2024 |

VII. CONCLUSION

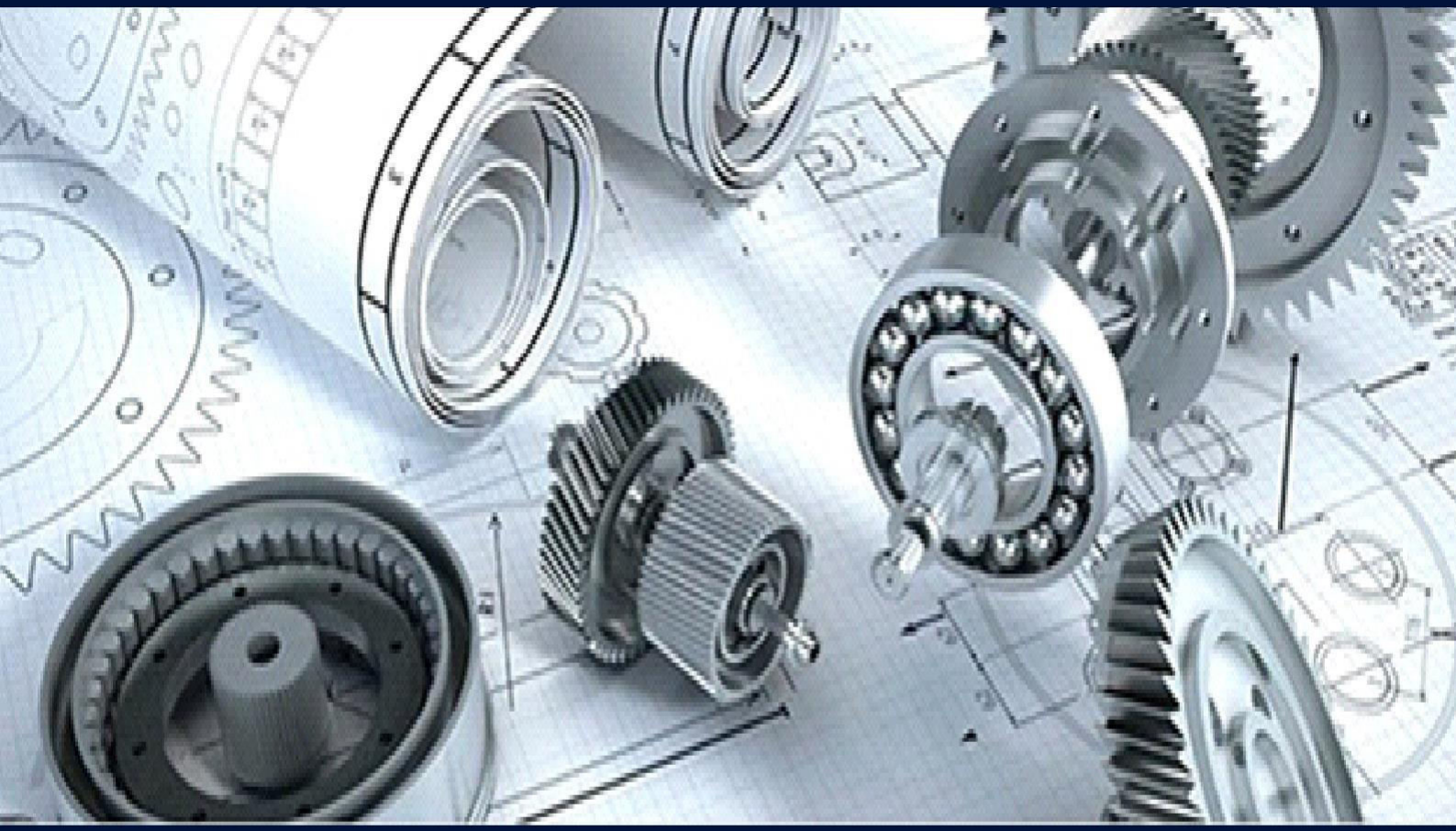
I have chosen from Kaggle the necessary characteristics and parameters for the used car prices dataset. Data and notebooks for data scientists and analysts are available on the open-source Kaggle platform for machine learning and data science. Using machine learning techniques, the necessary data is cleaned and preprocessed before any price prediction algorithms are applied. Next, we must apply train test split to separate the data into two portions for training and validation using train and test data, respectively, after pre-processing and cleaning the data. We must use a basic linear regression model to forecast the output and assess the correctness of the test and train. We must use a multiple linear regression model to train, test, and evaluate its accuracy. Linear Regression accuracy score is 0.90.

VIII. FUTURE SCOPE

In the future scope of this research on machine learning car price prediction using the Quikr dataset using linear regression, several enhancements and explorations can be pursued. Integrating more diverse and comprehensive datasets, including features like vehicle condition, historical pricing trends, and geographic factors, would provide a richer analysis. Exploring ensemble methods to combine multiple models might yield more robust predictions. Finally, developing an interactive application for real-time price prediction and continuously updating the model with new data could significantly enhance practical utility and relevance in the dynamic market of car sales.

REFERENCES

1. Oprea, C, Making the decision on buying second-hand car market using data mining techniques (Special, 2010), pp.17-26.
2. C Ozgur, Z Hughes, G Rogers and S Parveen, Multiple Linear Regression Applications Automobile Pricing (International Journal of Mathematics and Statistics Invention, 2016), pp.01-10
3. G.Chandrashekar and F. Sahin, “A survey on feature selection methods,” Computers Electrical Engineering, vol. 40, no. 1, pp. 16–28, 2014. [Online]. Available:http://www.sciencedirect.com/science/article/pii/S0045790613003066
4. Quikr_Car.csv data set from Kaggle.com
5. Nitis Monburinon, Suwat Rungpheung, Sabir Buya, Pitchayakit Boonpou, “Prediction of Prices for Used Car by using Regression Models” (ICBIR 2018)



INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING, TECHNOLOGY AND MANAGEMENT

+91 99405 72462 +91 63819 07438 ijmrsetm@gmail.com