

NIRMA UNIVERSITY

INSTITUTE OF TECHNOLOGY

**Machine Learning Project**

**On**

**“Car Price Prediction”**

**B. Tech CSE 2023**

Prepared by: -

Diksha Sharma 21bce058

Nupur Gandhi 21bce074

**Table of Contents**

1. Introduction

* Models Used
* Working of the code

1. Data Overview
2. Data Preprocessing
   * Exploring the Descriptive Statistics
   * Drop features that are not required
   * Missing value in dataset and its treatment
3. Data Exploration

* Exploring the probability distribution function (PDF)
* Dealing with outliers
* Checking the linearity
* Transform dependent variable using log-transformation
* Checking the multicollinearity using VIF

1. Feature Selection
   * LabelEncoder
   * Heatmap
   * Feature Importance
   * Categorical Features
2. Model Development
   * Declaring dependent and independent Variables
   * Feature Scaling
   * Train and Test Data Split
3. Linear Regression Model
4. Random Forest Regression Model
5. Gradient Boosting Regression Model
6. Best Model Selection
7. Manually Checking Predictions
8. Conclusion
9. Future Work
10. **INTRODUCTION**

To be able to predict used cars market value can help both buyers and sellers.

There are lots of individuals who are interested in knowing the price of used car in market at some points in their life because they wanted to sell their car or buy a used car.

In this process, it’s a big corner to pay too much or sell less then it’s market value.

In this Project, we are going to predict the Price of Used Cars using various features like Present\_Price, Selling\_Price, Kms\_Driven, Fuel\_Type, Engine Volume, Year, Mileage present in the dataset.

Thus, this report provides a detailed explanation of the entire project, including data preprocessing, exploratory data analysis (EDA), feature selection, model development, and evaluation of regression models like Linear Regression, Random Forest Regression and Gradient Boosting Regression. The best-performing model for car price prediction is identified, and its performance is analysed. Additionally, manual predictions and model selection are discussed.

* 1. **Models Used**

1. Linear Regression
2. Random Forest Regression
3. Gradient boosting regression
   1. **Working of the code**

We have written all the code in the .ipynb file.

To open it and run the code, open the .ipynb file on Jupyter notebook or google collab.

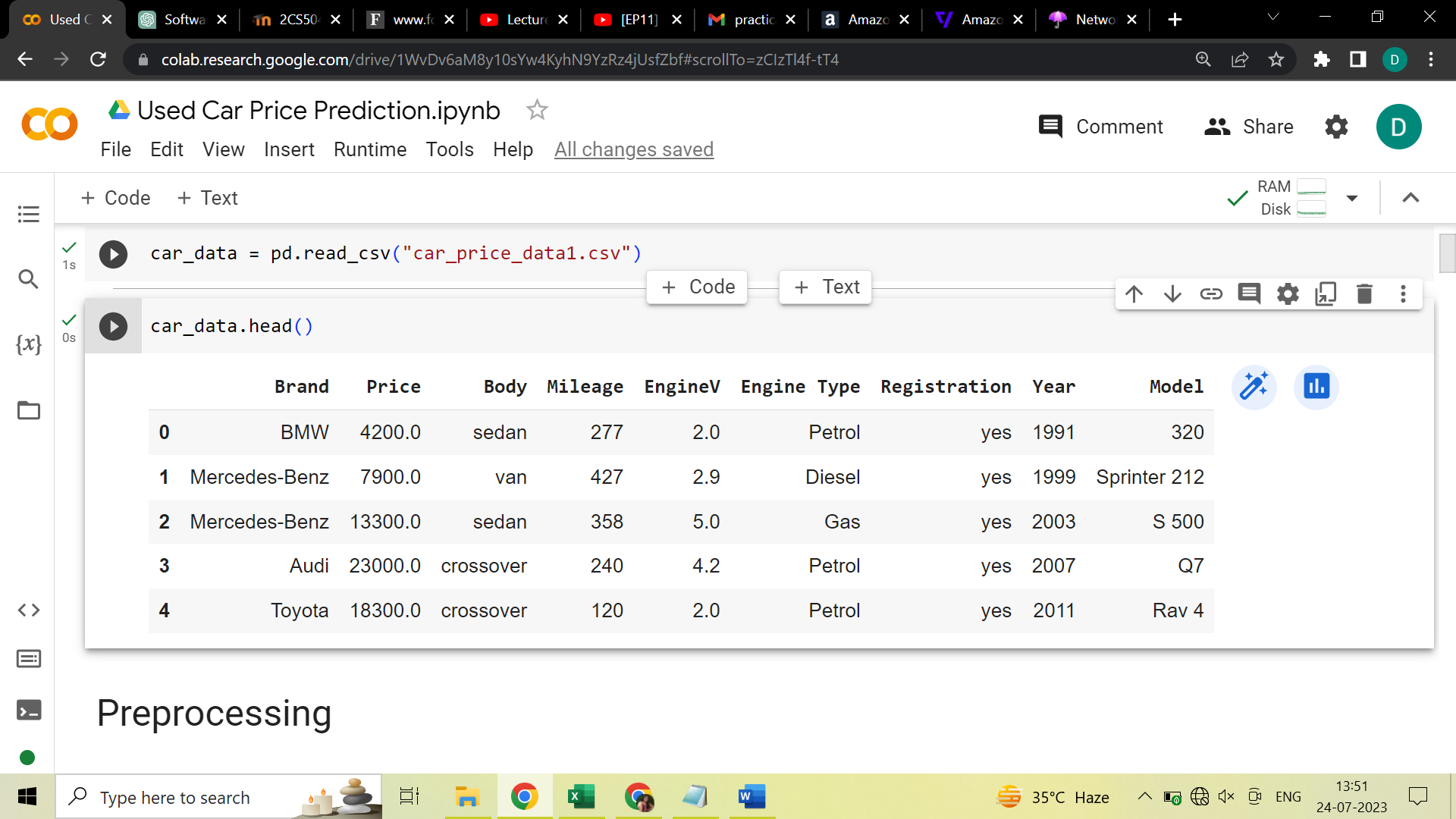
Before running the code, it is essential to install the various packages used in the prediction model.

* Numpy: - pip install numpy
* Matplotlib: - pip install matplotlib
* Seaborn: - pip install seaborn
* Pandas: - pip install pandas
* Sklearn: - pip install sklearn

Moreover, the dataset used in the prediction model has been uploaded. Before running the ipynb file, make sure to keep both the dataset and code file in the same folder or write the full pathname of the dataset in the code before running it. Else it will generate an error.

1. **DATA OVERVIEW**

This Project utilises a dataset containing information about various cars, including their prices and relevant attributes. Let’s load the dataset and have a quick look at its structure.



The dataset contains information about car prices and various features such as mileage, year of manufacture, engine volume, and model. The next step is to preprocess the data to make it suitable for model development.

1. **DATA PREPROCESSING**

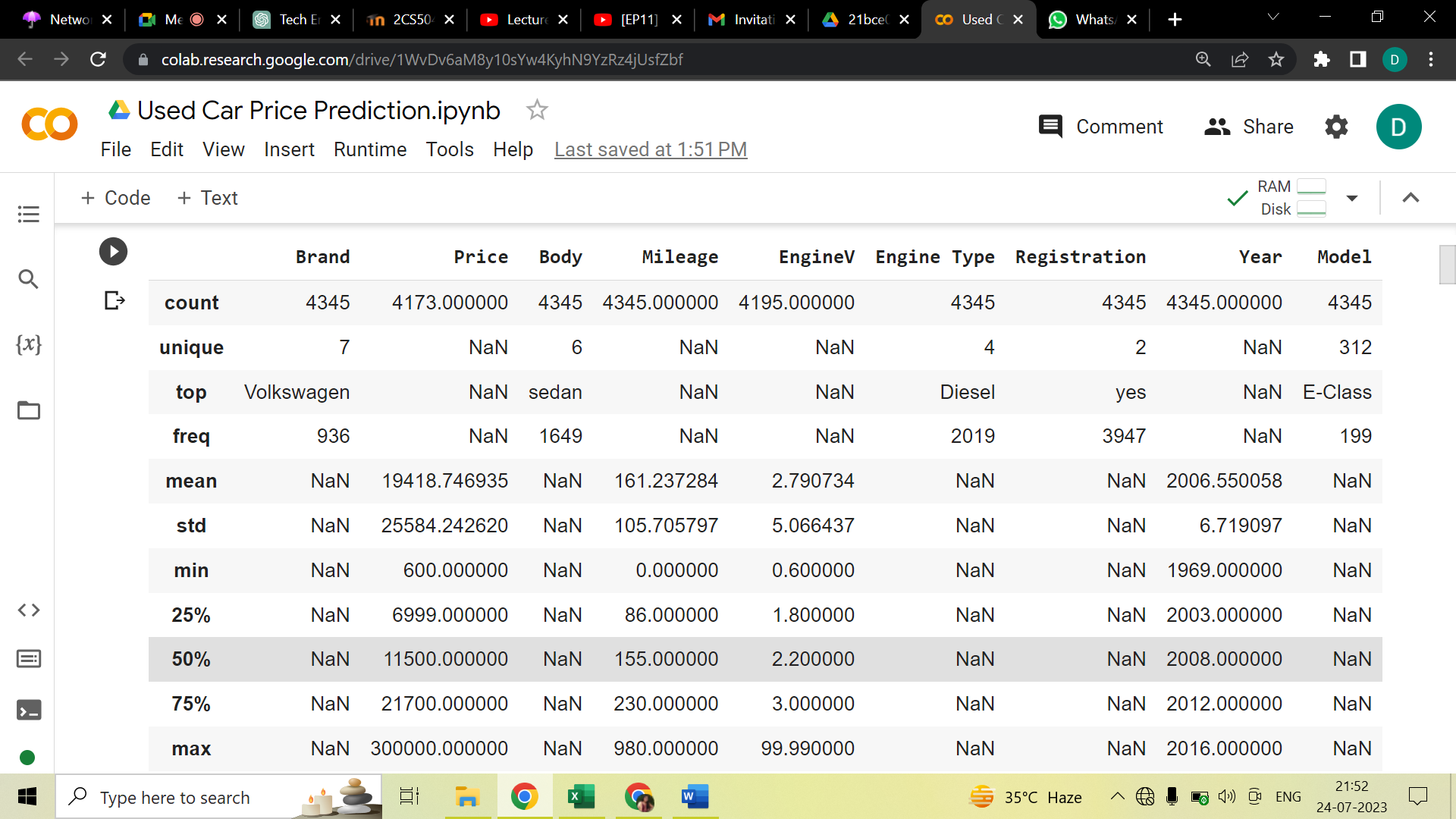
So, our first step was Data Preprocessing i.e. analysing the dataset on how it is present, the variables and its description.

For that, first we imported the necessary packages and modules required. Then we loaded the dataset.

Then we did the following steps: -

* 1. **Exploring the Descriptive Statistics of the variables.**

Descriptive statistics are computed for the dataset to gain insights into the data’s central tendencies and distribution. This step provides an understanding of the range and distribution of each feature.



* 1. **Drop features that are not required to build our model.**

We have car model column in our dataset which was not required to build the model, thus we dropped them.

* 1. **Check for any missing value in data set and treat it.**

We checked for any missing values present in any of the columns of our dataset and removed them lest they cause issues in our regression models.

Price and Engine volume column have null values so we have removed those values.

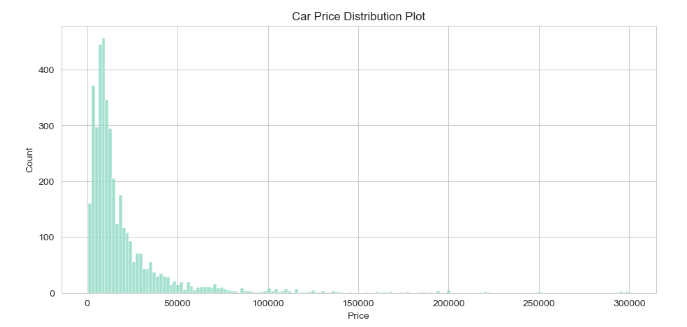
1. **DATA EXPLORATION**

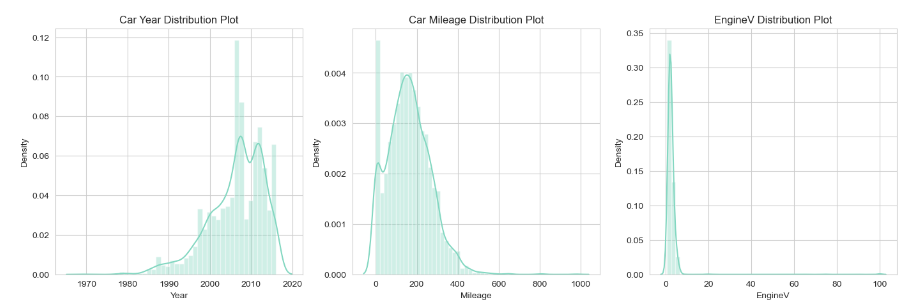
Our next step is analysing how each variable in the dataset is present, spotting the anomalies.

For that the following steps was followed: -

* 1. **Exploring the probability distribution function (PDF): -**

Visualizations, including histograms and scatter plots, are created to explore the distributions and relationships between the target variable (Price) and the independent variables. This step helps identify patterns and gain insights into the data.





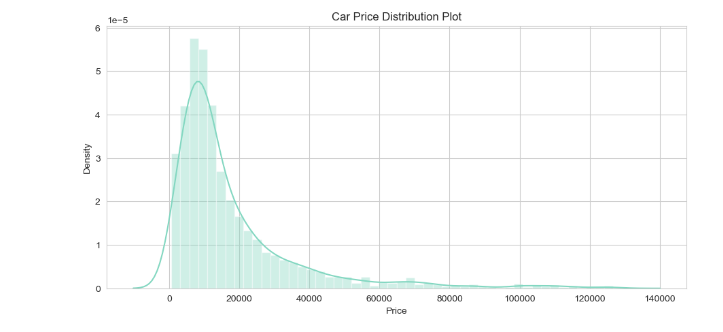
The distribution function graph clearly shows us some outliers present in our concerned variables. So, our next step is dealing with them.

* 1. **Dealing with the Outliers:**

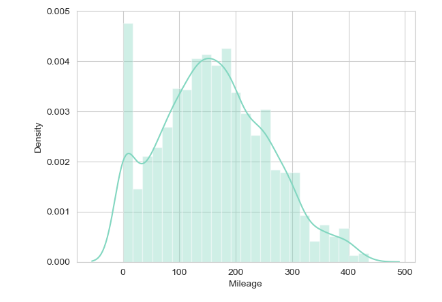
Outliers in the 'Price', 'Mileage', 'Year', and 'EngineV' columns are identified and handled using appropriate techniques such as truncation and transformation.

Here, we had remove the outliers from our dataset by truncating the extreme values.

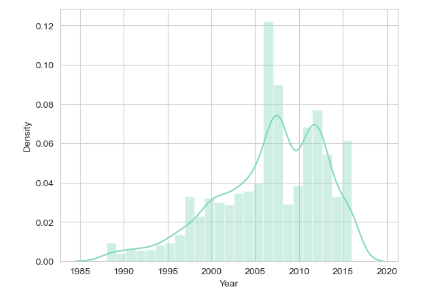
After removing the outliers, the graph of various features looks like this:-



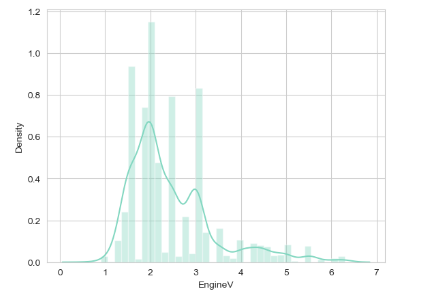
This is the graph of price and its density.



This is the graph of mileage distribution.



This is the graph of year density.



This is the distribution plot of EngineV distribution.

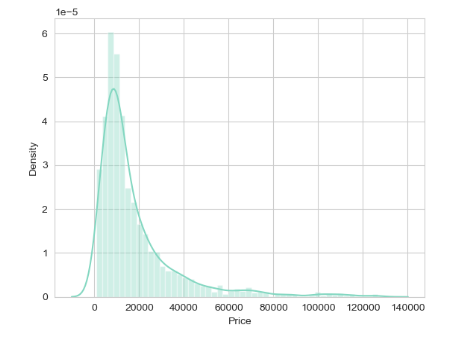
The updated distribution plot after removing outliers shows a better distribution of car prices.

* 1. **Checking the linearity:**

It shows how the dependent and independent variables are related and thus we can decide which regression model is suitable to fit in our dataset. It shows that our dependent variable is not linearly related to the independent variables thus to fit in the linear regression model, we need to transform the dependent variable.

We checked the linearity of our cleaned data using scatter plot.

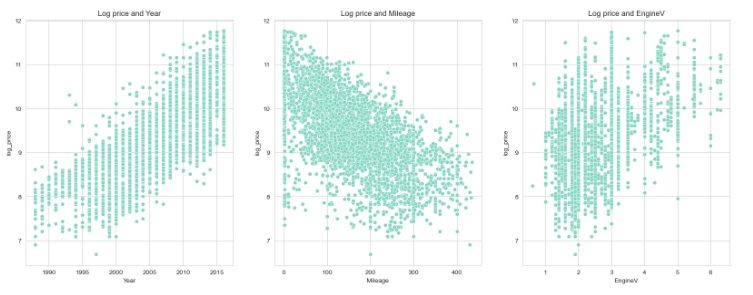


  
This shows that there is no linearity between the price and other features i.e., mileage, EngineV, Year.

* 1. **Transform independent variable using a log-transformation:**

We transformed our independent variable (**PRICE**) using log to make it linearly dependent with the dependent variables. Then we dropped the previous independent variable and added the one after transforming to our cleaned data.

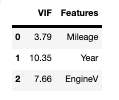
This shows the plot log(price) and year, mileage and EngineV.



* 1. **Checking the multicollinearity using VIF:**

We will check our variables for any possibilities of multicollinearity using the feature of VIF and fix them.

Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model. This can adversely affect the regression results. A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis.



Here, **Year** shows the highest VIF thus it is removed.

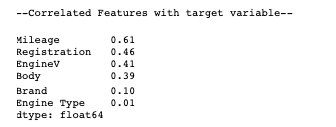
1. **FEATURE SELECTION**

The goal of feature selection techniques in machine learning is to find the best set of features that allows one to build optimized models of studied phenomena. It simplifies models, improves speed and prevent a series of unwanted issues arising from having many features

In this section, we followed the following steps: -

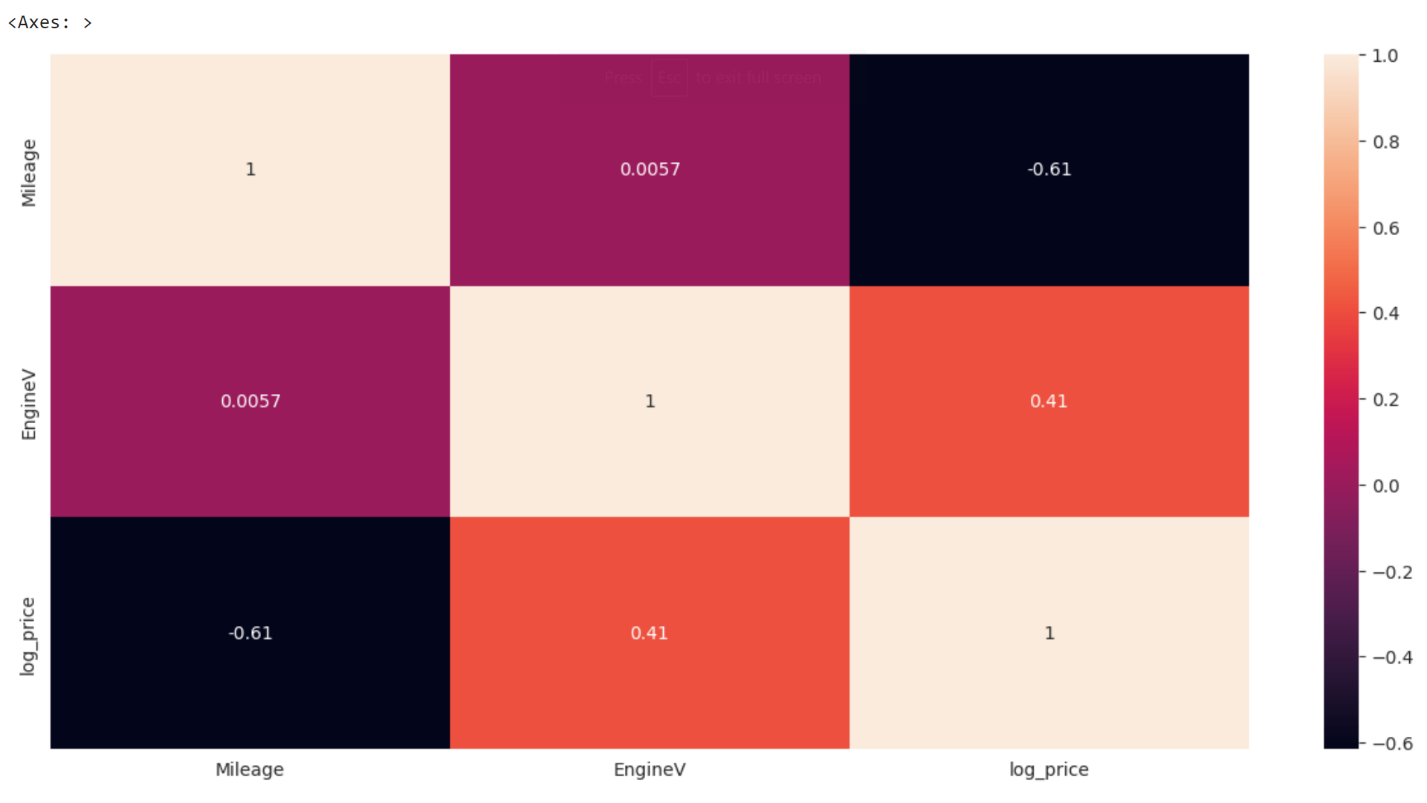
* 1. **Label Encoder(): -**

We first fit our dataset into label encoder function column by column and then checked the features correlated with the target variable.



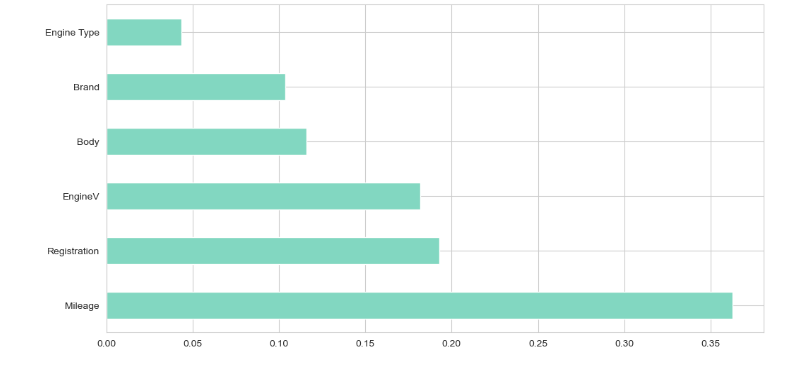
* 1. **Heatmap: -**

Then we generated a heatmap of our cleaned dataset. A heatmap is basically a graphical representation where individual values of a matrix are represented as colours. It is very useful in visualizing the concentration of values between two dimensions of a matrix. This helps in finding patterns and gives a perspective of depth.



* 1. **Feature Importance: -**

Feature importance gives a score for each feature of the data, the higher the score more important or relevant is the feature towards the Target variable. We used ExtraReggressor model to fit our dataset into it so as to find which feature is more dominant towards our target variable. Then we used a horizontal bar graph for its visualisation and sorted the values in descending order.



* 1. **Categorical features: -**

Our next step was converting the columns with categorial data i.e., non-numbers into numerical form so as to fit into our regression model. We used get\_dummies function which assigned a number to each different types of values present in the dataset column and then converted that row into multiple columns with values consisting of 0 for that property to not be relevant and 1 for it to be present.

1. **MODEL DEVELOPMENT:**

After the selection of important and relevant features our next step is developing the dataset according to our model needs.

The process is as follows:

* 1. **Declaring dependent and independent variables: -**

We decided upon the variables i.e., price is the dependent variable and all the other columns present in the cleaned dataset are independent variables.

* 1. **Feature scaling: -**

In this step we standardised the range of our independent variables. Our dataset contained features that are varying in degrees of magnitude, range and units.

Therefore, in order for our regression models to interpret these features on the same scale, we performed feature scaling.

* 1. **Train and test data:**

Next, we split our dataset into train and test datasets randomly.

Training dataset is used to train our model i.e., fit the model. Then we use test dataset to evaluate our dataset whether it can generalize well to the new or unseen dataset i.e., test set.

1. **LINEAR REGRESSION:**

First, we are using linear regression model.

Linear regression analysis is basically used to predict the value of a variable based on the value of another variable.

First, we fit our training dataset into the model and then predicted the values of the dependent variable i.e., the price of the car by using it on test dataset.

Then we found the

a) r2 score = 0.7726984972665855

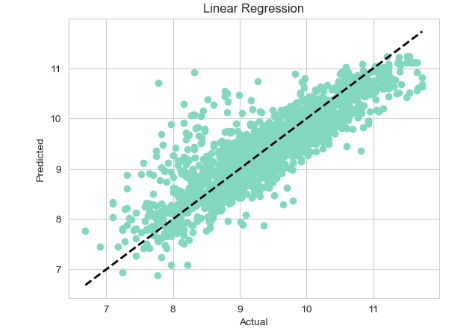
b) mean squared error = 0.3157034379634969

c) mean absolute error = 0.1847457945674902

**d) Root mean squared error = 0.4298206539563801**

of our predicted model to determine whether this regression is working well on it or not.

These results were found at default parameters.



1. **RANDOM FOREST REGRESSION:**

Next, we are using random forest regression model.

Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression.

It is a commonly-used machine learning algorithm which combines the output of multiple decision trees to reach a single result.

First, we fit our training dataset into the model and then predicted the values of the dependent variable i.e., the price of the car by using it on test dataset.

Then we found the

a) r2 score = 0.8127130881634755

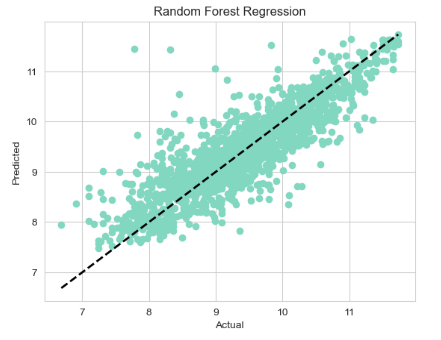
b) mean squared error = 0.27927374245340797

c) mean absolute error = 0.15222279185681664

**d) Root mean squared error = 0.390157393697488**

of our predicted model to determine whether this regression is working well on it or not.

These results were found at default parameters.



1. **GRADIENT BOOSTING REGRESSION:**

And lastly, we are using gradient booster regression model.

Gradient boosting Regression calculates the difference between the current prediction and the known correct target value.

This difference is called residual. After that Gradient boosting Regression trains a weak model that maps features to that residual.

Gradient boosting gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees.

First, we fit our training dataset into the model and then predicted the values of the dependent variable i.e., the price of the car by using it on test dataset.

Then we found the

a) r2 score = 0.812721686683953

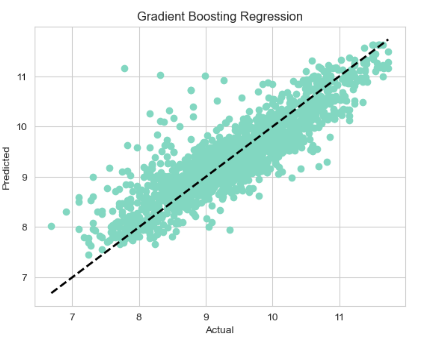
b) mean squared error =0.28799288828712005

c) mean absolute error = 0.1522158031634794

**d) Root mean squared error = 0.3901484373459407**

of our predicted model to determine whether this regression is working well on it or not.

These results were found at default parameters.



1. **BEST MODEL**

After working on all the three models. our next step is finding the model most suitable for our dataset.

For this we plotted the graph of all the three regression models i.e., linear, random forest, gradient boosting.

We plotted a scatter plot of actual vs predicted price.

And in analysing the plots, we found that gradient booster is the most suitable for our dataset as its scatter points are the most concentrated along the linear line among the three.

Thus, it produced the least error and gave us the most accurate predicted price.

RMSE found at RANDOM STATE=200 is as following

|  |  |  |  |
| --- | --- | --- | --- |
| Test Size | Estimators | Random forest regressor | Gradient Boost regressor |
| 0.2 | 100 | 0.393 | 0.430 |
|  | 200 | 0.394 | 0.418 |
|  | 300 | 0.393 | 0.412 |
|  | 400 | 0.392 | 0.410 |
|  | 500 | 0.392 | 0.409 |
|  | 700 | 0.393 | 0.409 |
|  | 900 | 0.393 | 0.410 |
|  | 1000 | 0.393 | 0.411 |
|  | 1500 | 0.393 | 0.416 |
|  | 2000 | 0.393 | 0.422 |
| 0.3 | 100 | 0.412 | 0.430 |
|  | 200 | 0.409 | 0.418 |
|  | 300 | 0.409 | 0.412 |
|  | 400 | 0.410 | 0.410 |
|  | 500 | 0.410 | 0.409 |
|  | 700 | 0.410 | 0.409 |
|  | 900 | 0.410 | 0.410 |
|  | 1000 | 0.410 | 0.411 |
|  | 1500 | 0.410 | 0.416 |
|  | 2000 | 0.410 | 0.422 |
| 0.5 | 100 | 0.435 | 0.427 |
|  | 200 | 0.435 | 0.419 |
|  | 300 | 0.434 | 0.418 |
|  | 400 | 0.435 | 0.417 |
|  | 500 | 0.435 | 0.418 |
|  | 700 | 0.434 | 0.422 |
|  | 900 | 0.434 | 0.427 |
|  | 1000 | 0.434 | 0.429 |
|  | 1500 | 0.434 | 0.438 |
|  | 2000 | 0.435 | 0.445 |

At different values of the random state,

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Random State | Test Size | Linear Regression | Random Forest Regression | Gradient Booster Regression |
| 100 | 0.1 | 0.434 | 0.428 | 0.410 |
|  | 0.2 | 0.434 | 0.408 | 0.409 |
|  | 0.3 | 0.428 | 0.413 | 0.404 |
|  | 0.5 | 0.438 | 0.424 | 0.414 |
| 200 | 0.1 | 0.436 | 0.407 | 0.411 |
|  | 0.2 | 0.440 | 0.392 | 0.412 |
|  | 0.3 | 0.453 | 0.414 | 0.430 |
|  | 0.5 | 0.444 | 0.437 | 0.427 |
| **365** | **0.1** | **0.403** | **0.376** | **0.361** |
|  | 0.2 | 0.429 | 0.390 | 0.390 |
|  | 0.3 | 0.450 | 0.416 | 0.407 |
|  | 0.5 | 0.450 | 0.432 | 0.417 |

This table indicates that the best model is Gradient Boosting regression model with **RMSE = 0.361** at a **random state = 365** and **test size = 0.1** with default estimators.

1. **MANUALLY CHECKING THE PREDICTIONS**

For manually checking the error in our predictions, we followed the following steps.

First, to find the actual price, we took the exponential of the price column present in our dataset as we had log transformation on it during preparation of data.

Then, we found the residual i.e., difference between the actual and predicted price and residual in percentage.

Then, we plotted it in the form of table to show clear results of our models.



1. **CHECKING OUR MODEL**

And lastly, we took one row from our cleaned dataset and predicted it's price by directly fitting the row values.

1. **CONCLUSION**

The car price prediction machine learning project successfully developed and evaluated regression models for predicting car prices. After analysing the performance of three regression models(Linear Regression, Random Forest Regression, and Gradient Boost Regression), we determined that the Gradient Boost Regression model is the most suitable for car price prediction. It exhibited the least error and provided the most accurate predicted prices. The model can be utilized to estimate car prices based on relevant attributes, making it a valuable tool for both buyers and sellers in the used car market.

By employing data preprocessing techniques, exploratory data analysis, feature selection and regression model evaluation, we achieved a robust and accurate car price prediction model. The insights gained from this project can be applied to similar datasets to develop predictive models in other domains as well. The machine learning models created in this project can be further improved and fine-tuned by incorporating more data and optimising hyperparameters, leading to even more accurate car price predictions.

1. **FUTURE WORK**

The car price prediction project can be further improved and extended in the future. Some of the potential future works includes:

* Collecting more data to improve model accuracy
* Experimenting with different regression algorithms or ensemble methods.
* Feature engineering to create new relevant features.
* Deploying the model as a web application for real time car price prediction.

Overall, this project demonstrates the application of machine learning in predicting car prices and highlights the importance of importance of model evaluation and selection to achieve accurate and reliable predictions.