

CAPSTONE REPORT

*USED CAR PRICE PREDICTION*

**Submitted to:**

Concerned Faculty

At

Great Learning

**Submitted by:**

Anshuman kumar

ER Sumanth

Elumalai .M

Sravani .B

PGPDSE-FT Bangalore October 2023

Post Graduate Program in Data Science and Engineering

**CONTENTS**

Page No.

|  |  |
| --- | --- |
| 1.Introduction to the business problem | 3 |
| 2. Exploratory Data Analysis | 4 – 18 |
| 3. Business Insights from EDA | 19 |
|  |  |
|  |  |

**Data dictionary :-**

* id (int): Record id
* region (Object): Region where the car is being sold
* price (float): Selling price of the car
* year (Object): Manufacturing Year of the car
* manufacturer (Object): Car Manufacturer name
* model (Object): Car Model name
* condition (Object): Condition of the used car (like: Excellent, New, Good etc.,)
* cylinders (Object): Number of cylinders in the engine
* fuel (Object): Type of fuel used in the car
* odometer (float): Number of miles driven
* title\_status (Object): Choice of usage (like: Clean, Salvage, Rebuilt etc.,)
* transmission (Object): Indicates the transmission type (like: Manual, Automatic, others)
* VIN (Object): VIN Number
* drive (Object): Indicates the type of drive (like: fwd, rwd, 4wd)
* size (Object): Indicates the size of the car (like: full-size, mid-size, compact, sub-compact)
* type (Object): Indicates the vehicle segment (like: SUV, Coupe, Convertible etc.,)
* paint\_color (Object): Car color
* description (Object): Remarks by the seller on the car condition
* county (Object): County where car is being sold
* state (Object): State where car is being sold
* lat (object): Latitude of the selling place
* long (object): Longitude of the selling place
* posting\_date (DateTime):Car sale posting date

**Introduction:**

We aim is to showcase a few of the basics of data science (data cleaning, encoding, feature engineering, and model training), all while attempting to solve a problem that is common among businesses to satisfy the customer’s expectation.

we will attempt to create a model that predicts whether a customer was **satisfied** or **unsatisfied** with the experience and/or service which an airline provided.

The below points include the necessary things from a buyer perspective:

**Problem Statement:**

The used car market suffers from inconsistent data quality and a lack of predictive models for fair pricing, hindering informed decision-making.

This project aims to leverage multiple machine learning models to enhance data quality, predict fair market values accurately, and provide user-friendly interfaces for analysis.

By integrating various algorithms, the solution seeks to offer comprehensive insights and facilitate efficient decision-making in the used car industry.

**Objective:**

Take advantage of all of the feature variables available below, use it to Analyze and predict used car price.

# Exploratory Data Analysis:

**Solution:**

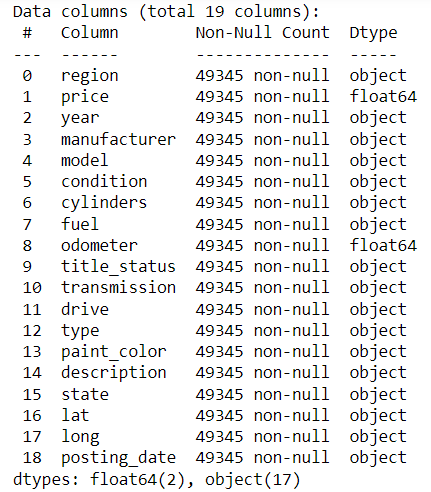


Firstly, after importing all the relevant libraries on Jupyter notebook, we load the data set. Then, we perform EDA to extract and see patterns in the given data set.

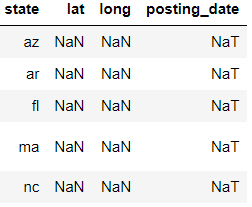
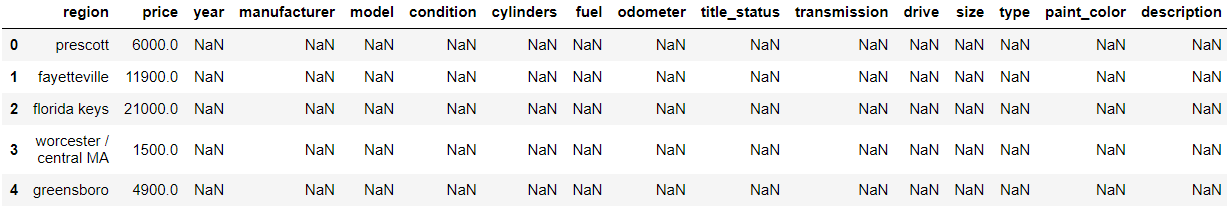
**Shape of the data:**

The shape of the raw data was ((50000, 23) , after performing EDA that is the data consists of 49346 rows 19 columns.

**Data types of the Data:**

****

**First 5 records in the data:**

****



# 

**Null values:**



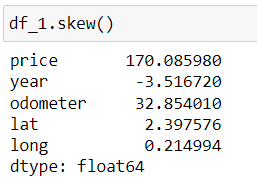
The Null values that are present in the dataset:

# 

# Duplicates:

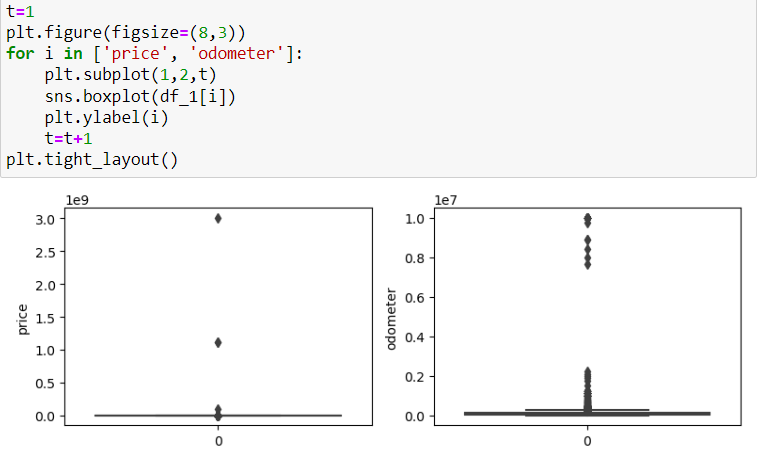
There are no duplicates present in the dataset.

**Skewness of the dataset:**

****

**Boxplot:**

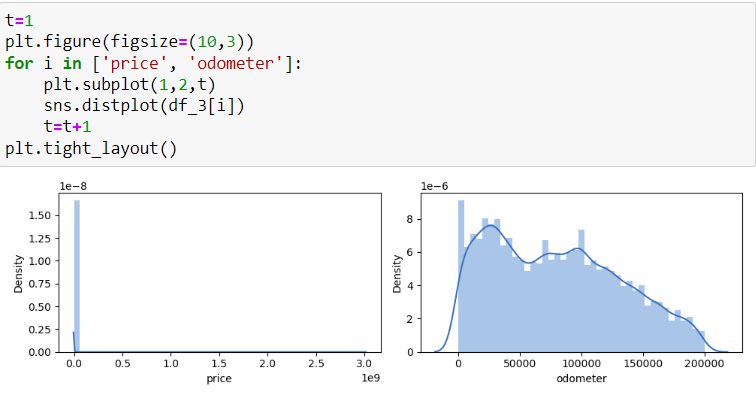
**Outliers present in the dataset:**

****

# OUTLIER TREATMENT:-

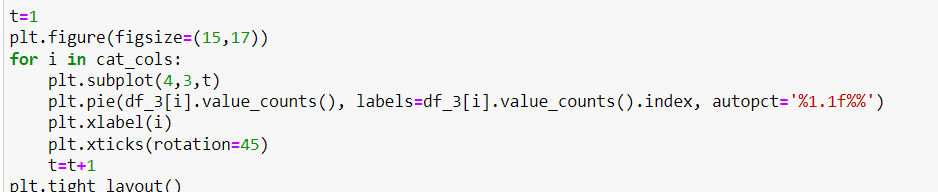
1. We cannot ignore outliers for price variable, since vehicles with higher prices bring more revenue and profits to the company.
2. Assuming there is 10% profit on selling price, we consider vehicles only which has a profit of minimum 50 USD (implies selling price = 500 USD).
3. We can remove right outliers from odometer variable, since we rarely get vehicles that have been driven for exceptionally long distance.
4. We dont want to remove left outliers for odometer variable since price & profits by selling almost new vehicle will be high and the vehicles will be fast selling.

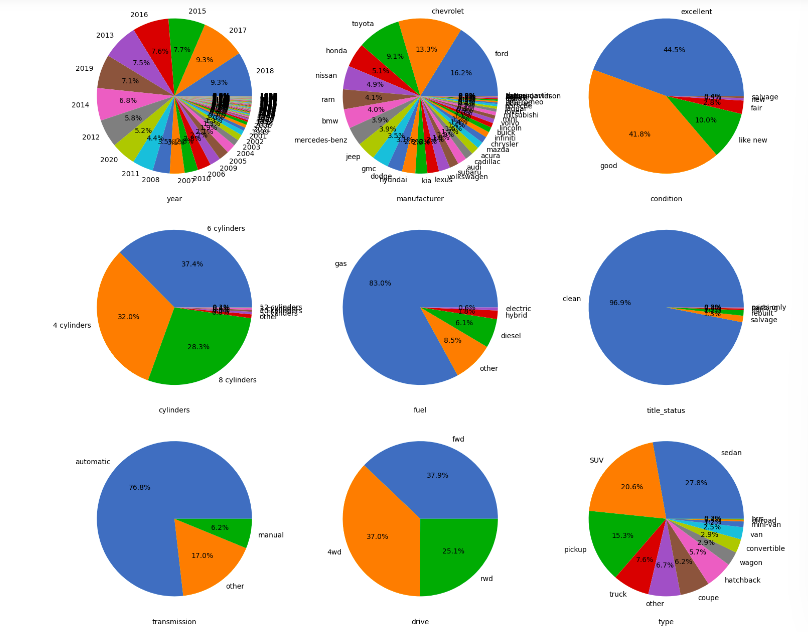
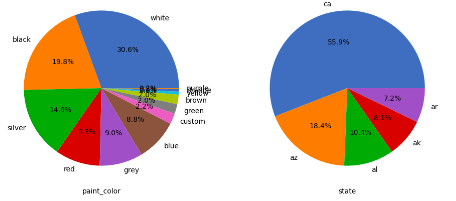
**Uni-Variate Analysis:**

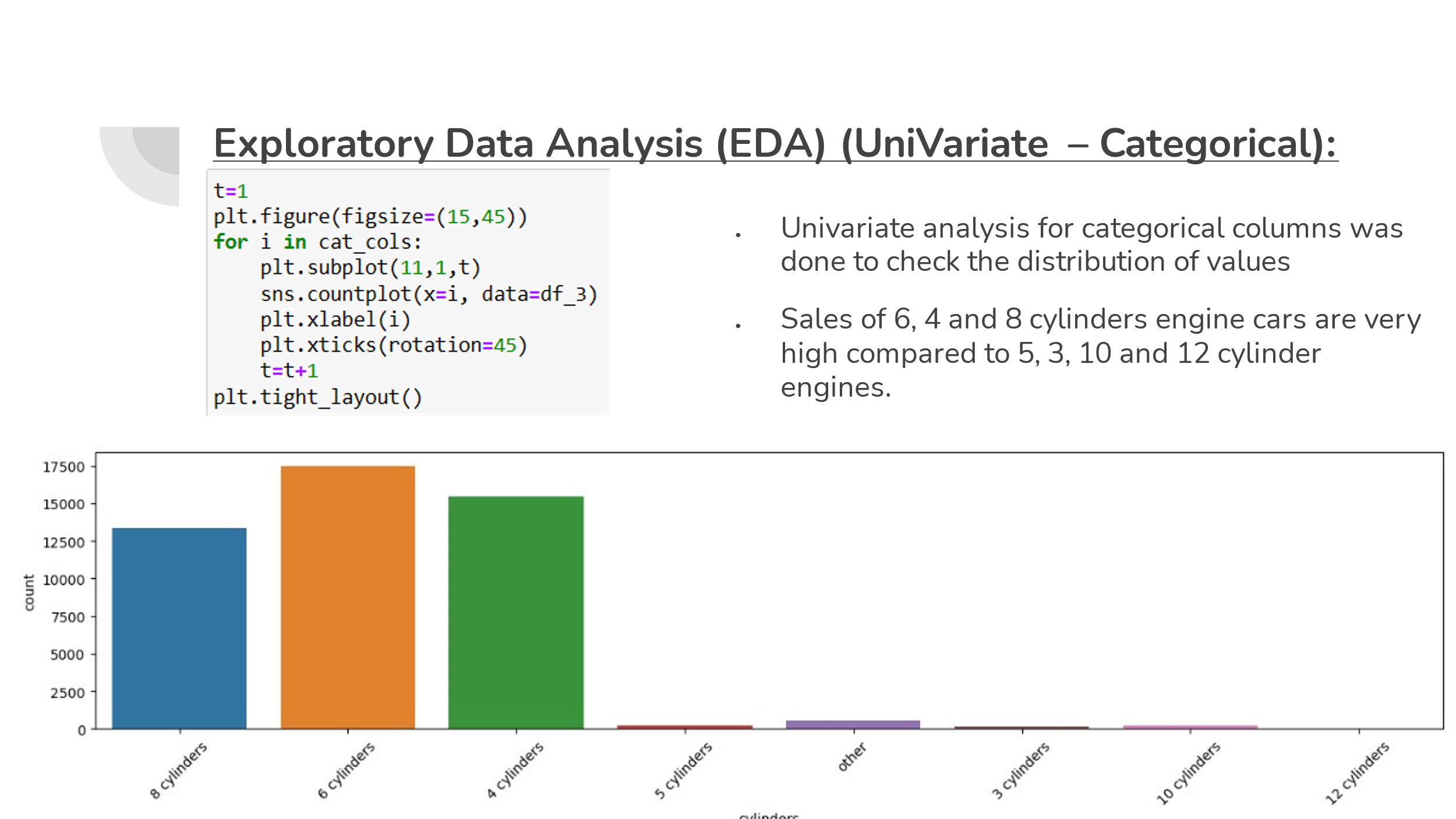


Univariant for numerical columns of Price and odometer.

1. **Representing the percentage distribution of categorical columns** :





# 

# Bivariate Analysis:

# Distribution of each service based on categorical

# 

# 

# 



* **Numeric Transformation:** Ordinal encoding converts categorical variables into numerical representations (e.g., manufacturer, condition).
* **Sequential Value Assignment:** Each category within a variable is assigned a unique numerical value starting from 0 (e.g., cylinders, fuel type).
* **No Inherent Hierarchy:** Ordinal encoding doesn't imply hierarchy among categories (e.g., manufacturer, fuel type).
* **Suitability for Ordinal Data:** Effective for variables with predefined order (e.g., condition, title status).
* **Enhanced Model Performance:** Contributes to improved predictive accuracy in tasks such as estimating used car price.

# Splitting the data:

# Split the dataset into training and testing sets. This allows you to train the model on one set and evaluate its performance on another independent set.

# Base Model:

# Linear regression offers simplicity, interpretability, and efficiency, making it a suitable choice as a base model for predicting.

# 

# We have a high MSE value (around 60% the mean value)

# Though R-Squared (Explained Variance) is less but the MAPE value is good

# Standard Errors assume that the covariance matrix of the errors is correctly specified.

# The condition number is large, 3.2e+05. This might indicate that there are strong multicollinearity or other numerical problems.

# We need to improve the model performance by changing outlier treatment, transformation, encoding, model selection.

# *Base model before removing outliers on right side.*

# 

# 

# Since R-Squared is very low, we need to further enhance model performance by treating multicollinearity.

# removing outliers from price which are less than the lower limit and outliers from odometer greater than the upper limit.

# removing outliers from odometer variable, since vehicles driven for exceptionally long distance are rare.

# outliers for price variable can’t be ignored, vehicles with higher prices bring more revenue and profits.

# For the sake of parameter influence we shall remove outliers (df\_3) for basic model building

# Rebuild the model including outliers of price (df\_2) and improve the efficiency

# Standard Errors assume that the covariance matrix of the errors is correctly specified. The smallest eigenvalue is 4.53e-31. This might indicate that there are strong multicollinearity problems or that the design matrix is singular.

# 

# Base model after removing the outliers on both sides:

# 

# Since R-Squared is very low, we need to further enhance model performance by treating multicollinearity.

# High VIF indicating variance of the coefficient estimate is greatly inflated by multicollinearity.

# Independent variable is highly correlated with other independent variables in the model.

# Natural Language Processing to car dataset :

# Learned extracting insights from textual descriptions.

# Enriching our feature set with additional context.

# This iterative process has enhanced our skills in data analysis and predictive modeling.

# 

# Decision Tree :

# Building a decision tree model to check how well the model performs on our train and test datasets. A non-parametric, interpretable model is used for regression tasks for building a tree structure based on feature thresholds in this case Car\_Type to predict the target variable.

# 

# 

# As we can see, we obtained a test MSE score of test 8273 and train RMSE of 8093.

# In this case the premium car which is older than the economy car could be priced same .

# So Decisiontreee helps us making nodes segment wise decision tree making them more robust to outliers.

# XGB-Regressor : - Before implementing NLP :-

# 

# XGBoost's boosting key properties in this car dataset :-

# Enabling algorithm to iteratively improve predictions.

# Focusing on areas where previous models struggled.

# Effectively capturing the complex relationships between car features and sales prices.

# Boosts predictive accuracy.

# XGBoost (Extreme Gradient Boosting)

# Is well-suited for our used car dataset. It efficiently handled complex relationships between car features (like year, mileage, condition, etc.) and sales prices.

# Ability to handle missing data and provide feature importance scores,

# XGBoost can effectively predict car sales prices based on the provided features, offering accurate insights for our regression task.

# Using sequential feature selector to obtain the features that most significant using GridSearchCV.

# This ensures that XGBoost is optimized for our specific regression task, improving its performance in estimating used car prices effectively.

# GridSearchCV for this dataset :-

# Helps fine-tune XGBoost's parameters

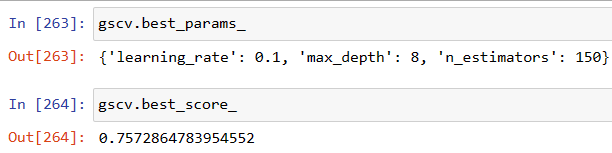
# Learning rate

# Maximum tree depth regularization parameters

# Enhancing the model's predictive accuracy

# Robustness.

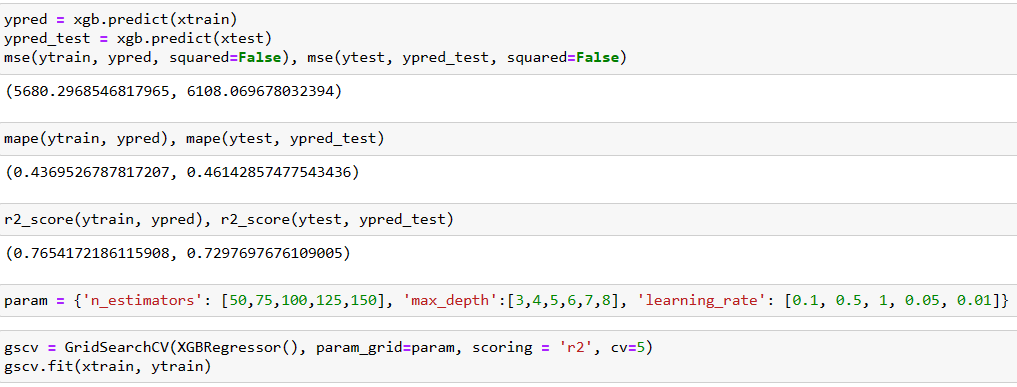




This ensures that XGBoost is optimized for our specific regression task, improving its performance in estimating used car prices effectively.

**XGB-Regressor : - After implementing NLP :**





# The lack of improvement in scores suggests that the textual information is not sufficiently informative for predicting used car prices.

# Despite the lack of improvement with NLP features, we have finalized XGBoost as our best model.

# We've achieved satisfactory results in terms of R-square, MSE, and MAPE values, indicating that XGBoost effectively captures the underlying.

# *Summary*

# Built a Robust predictive model for sales price for used cars.

# Concluded essential features like odometer, condition and cylinders.

# Using ML techniques found more critical features like manufacturer, model, fuel type.

# Significance improvement in predictive model performance due to critical features addition.

# Experimented xgboost , DecisionTree regressor , NLP to extract and enhance predictive accuracy.

# *Challenges:*

# Data cleaning

# Data encoding

# Binning

# Highly skewed data

# Very complex patterns.

# Feature relevance

# Model selection.

# *Learning outcomes*:-

# Intricacies of predictive modelling

# Feature engineering.

# Domain specific knowledge.

# Iteratively adaptive strategies for optimal performance.

# *Way forward* :-

# Collecting more informative features such as mileage , how many times the car is pre –owned for better feature relevance.

# Try other types of encoding and transformation techniques.

# *Inference :-*

# This dataset consists of inconsistent data quality and lack predictive modeling for fare pricing.

# of all the models that are build by the feature engineering XGB regressor is better in predicting fare price.