**Project name – Car Price Prediction Using Decision Tree Algorithm**

**Prepared by – Madhav Kamani**

**Business case :**

Predicting car prices using decision trees offers a compelling business case for various stakeholders in the automotive industry. Car dealerships can accurately price vehicles, attracting customers with fair deals while maintaining profitability. Buyers benefit from informed decisions, avoiding overpayment, and building trust with sellers. Private sellers optimize listing prices, speeding up sales. Manufacturers and analysts gain insights into market trends, aiding in designing new models and adjusting production. Online resale platforms integrate price predictions, enhancing user experience. Overall, decision tree-based car price predictions enhance transparency, customer satisfaction, and informed decision-making across the automotive ecosystem.

To be able to predict used cars market value can help both buyers and sellers. There are lots of individuals who are interested in the used car 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.

**Dataset Overview:**

In this project we will be working on Vehicle dataset. This dataset contains information about used cars. We are going to use for finding predictions of price with the use of Decision regression models.

The datasets consist of several independent variables include:

1. Car name:

It contains the name of the vehicle and since different clients bought the same model in various years , the names of the vehicles are repeated.

1. Year:

This feature contains the car’s manufactured year.

1. Selling price

The owner wants to sell the car to others in the present year 2023 (in lakhs).

1. Present price

Ex-showroom price (in lakhs).

1. Kms driven

The kilometers, the car has travelled till now.

1. Fuel type

It means the cars has 3 fuel types. They are Diesel, Petrol and CNG.

1. Seller type

In seller type, the car is selling by the individual owner and Dealer.

1. Transmission

It indicates the car has two gear types i.e., Manual and Automatic.

1. Owner

It tells us how many owners the car has in previous.

**Exploratory Data Analysis:**

The main purpose of EDA is to help look at data before making any assumptions. It can help us to identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.

The Exploratory Data Analysis (EDA) phase is crucial for understanding the underlying patterns, relationships, and potential challenges present in the car price prediction dataset. This section aims to provide a comprehensive overview of the dataset's main features and their relevance to car price prediction.

* **Missing data**

For various reasons, many real world datasets contain missing values, often encoded

as blanks, NaNs or other placeholders. Such datasets however are incompatible with

scikit-learn estimators which assume that all values in an array are numerical, and

that all have and hold meaning.

In Car price prediction dataset, there is no null values present in data.

* **Bar plot** :

A bar plot is a graphical representation that shows data values using rectangular bars. Every bar represents a class or group, and its height or length reflects the amount of data it represents. Bar plots are frequently used to display the distribution of a categorical variable or to compare various categories.

Bar plots make it simpler for clients to understand and evaluate the outcomes of your car price prediction analysis by visually summarizing and presenting important details about the data and the model's performance.

1. In the above plot, we can see that Selling Price of cars seems to have higher prices when sold by Dealers when compared to Individuals.

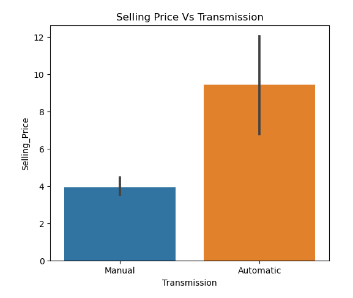


Fig 1 : Selling type vs Seller type

1. It can be observed that Selling Price would be higher for cars that are Automatic compared to Manual cars.

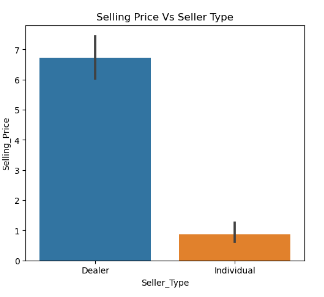


Fig 2 : Selling price vs Transmission

1. Selling Price of cars with Fuel Type of Diesel is higher than Petrol and CNG.

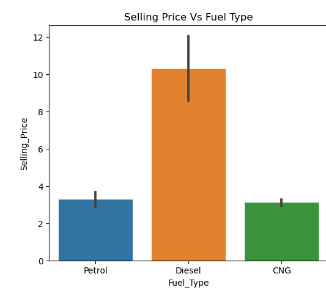


Fig 3 : Selling price vs Transmission

1. Selling Price is high with less Owners used Cars and low price for more owners used cars.

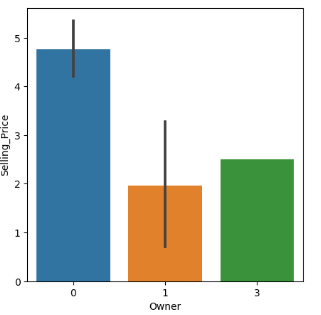


Fig 4 : Selling price vs owners

1. Selling Price of cars 2 years old would be high and gradually decreases with car of 17 years old.

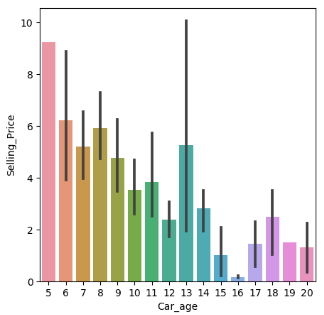


Fig 5 : Selling price vs car age

* **Heatmap** :

A heatmap is a graphical representation of data where individual values are represented as colors within a grid. Heatmaps are particularly useful when working with tabular data, as they provide a visual way to display relationships, patterns, and trends among variables in a dataset.

Heatmaps are often used to visualize the correlation between variables. Correlation indicates how two variables are related to each other. In a heatmap, colors are used to represent the strength and direction of the correlation between pairs of variables. Common color schemes include gradients of colors like red and blue, where darker shades represent stronger correlations (positive or negative) and lighter shades indicate weaker or no correlations.

In Car price prediction, heat map shows the correlation between variables. Where, Selling price is the target variable.

Feature engineering involves the process of creating, transforming, or selecting features to enhance the predictive power of machine learning models. In the context of car price prediction, careful feature engineering can uncover hidden patterns and relationships that contribute to accurate predictions.

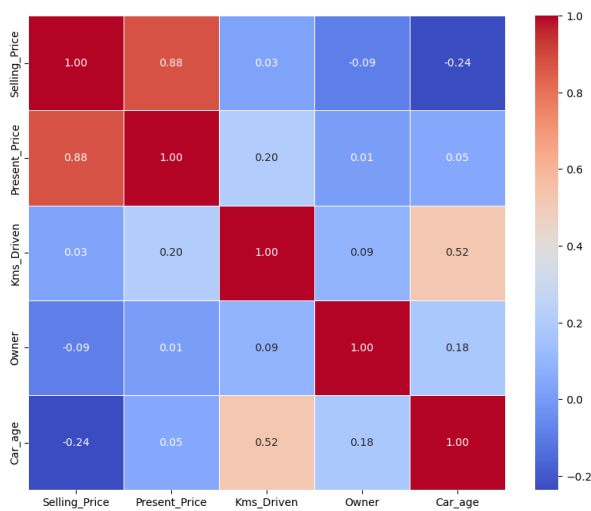
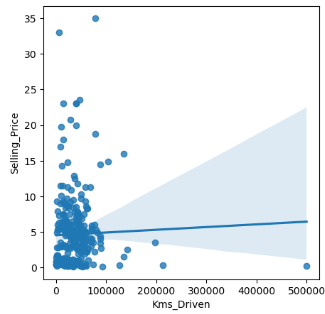


Fig 6 : Heatmap

* **Regplot (regression plot)**

creates a regression line between 2 parameters and helps to visualize their linear relationships.

1. Selling price vs kms driven : Lesser the Kms driven higher the Selling Price.

Fig 7 : selling price vs kms driven

1. Selling price vs present price : selling price is lesser than the present price(ex show room price.

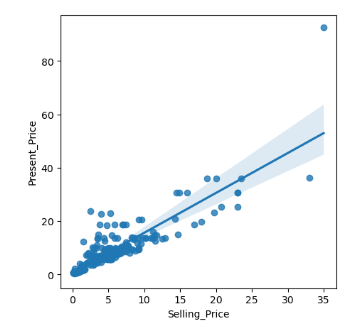


Fig 8 :selling price vs present price

**Feature Engineering:**

power of machine learning models. In the context of brain stroke detection, careful feature engineering can uncover hidden patterns and relationships that contribute to accuratepredictions.

* **Label Encoding:**

Label encoding is a technique used to convert categorical data into numerical data, allowing machine learning algorithms to process and analyze the data effectively. In label encoding, each unique category or label in a categorical variable is assigned a corresponding integer value.

In my data Fuel type, Seller type, Transmission are the categorical data.

So, converted those into numerical data by using label encoding.

cpp['Fuel\_Type'].replace({"Petrol": 1, "Diesel": 0, "CNG": 2}, inplace=True)

cpp['Seller\_Type'].replace({"Dealer": 1, "Individual": 0}, inplace=True)

cpp['Transmission'].replace({"Manual": 1, "Automatic": 0}, inplace=True)

**Model Building:**

1. We have split the data into train and test to train and test the model. Sklearn gives

these feature of splitting dataset using train\_test\_split. It will split the

data into X\_train, X\_test, y\_train and y\_test.

1. Then, import the model decision tree regressor, then fit the model into x\_train

and y\_train.

1. Then I did a hyperparameter tuning to train a model for better performance.
2. I have fitted x\_train and y\_train into decision tree regressor to train the model.
3. Then I have predict , whether the model is giving accurate values or not.
4. As my dependent variable is continuous .so, I have imported r2\_score to

check the accuracy

1. Then I check score for x\_train and y\_train to get bias value. Similarly , I checked

the score for x\_test and y\_test to get variance.

**Results:**

We have built the Decision tree model, where I got the results for accuracy , bias

and variance are

r2\_score : 0.8572640488749044

Bias : 0.8970544467935754

Variance : 0.8572640488749044

**Conclusion:**

In Decision Tree, We got accuracy 85% bias as 89% and variance as 85% is some what average accuracy but not upto the mark. So, I will conclude that for predicting a car price decision tree will give average accuracy.