Car Price Prediction

Submitted by

Steffin Varghese

Batch - 26

*In partial fulfilment of Data Science – Internship*

*At*



Flip Robo Technologies

AI and Software Development company

Flip Robo Technologies | Indiranagar, Bengaluru - 560 038, Karnataka, India

Month of Submission

July 2022

Acknowledgement

I am highly indebted to FlipRobo Technologies for giving me this opportunity to work on a project and for the guidance and persistent supervision, as well as for providing necessary project information and assistance in completing the project.

I would want to convey my gratitude to the members of FlipRobo Technologies for their kind encouragement and support in completing this project.

I would like to extend my heartfelt gratitude and appreciation to SME, Ms **Khushboo Garg** for spending such close attention to me and assisting me during the project's completion, as well as towards others who have volunteered to assist me with their skills.

MACHINE LEARNING FOR PROPERTY PRICE PREDICTION AND

PRICE VALUATION

I acknowledge my gratitude towards the authors of papers: “**Predicting the Price of Used Cars using Machine Learning Techniques**”, “**Used Car Price Prediction**” and “**Car Price Prediction Using Machine Learning”** for the insights I could attain through the extensive research which enhanced my knowledge in development of the project.

**CONTENTS**

**INTRODUCTION**

* Business Problem Framing
* Conceptual Background of the Domain Problem
* Review of Literature
* Motivation for the Problem Undertaken

**Analytical Problem Framing**

* Mathematical/ Analytical Modelling of the Problem
* Data Sources and their formats
* Data Pre-processing
* Data Inputs- Logic- Output Relationships
* Assumptions related to the problem under consideration
* Hardware and Software Requirements and Tools Used

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)
* Testing of Identified Approaches (Algorithms) and evaluation of selected models
* Key Metrics for success in solving problem under consideration
* Visualizations
* Interpretation of the Results

**CONCLUSION**

* Key Findings and Conclusions of the Study
* Learning Outcomes of the Study in respect of Data Science
* Limitations of this work and Scope for Future Work

**INTRODUCTION**

**Business Problem Framing**

With the covid 19 impact in the market, we have seen lot of changes in the car market. Now some cars are in demand hence making them costly and some are not in demand hence cheaper. One of our clients works with small traders, who sell used cars. With the change in market due to covid 19 impact, our client is facing problems with their previous car price valuation machine learning models. So, they are looking for new machine learning models from new data.

**Conceptual Background of the Domain Problem**

Used car selling market was always a booming market. With the rise to auto owners, the market for the used cars always increased. The main reason behind this is the introduction of new variants with new features and more efficiency cars to the market by the automakers. This will create a feeling in the mind of car owners of the car being obsolete, which makes them to sell the existing car and look for a new one. This cycle has been continuing and will be continued in the future as well. This is the main reason for the supply of cars in the used car market, as the seller wants to get rid of the used car to buy new one, most of them will be ready for a price negotiation to get the process done quickly. There are buyers who exploit this and get the car for much lower prices. So, there is always a demand and supply of cars in the used car market. Everything was running smoothly before the wide spread of Covid 19. After the pandemic, every market including the automobile industry became the victim and incurred a knock. Due to the inflation, economic instability, etc., the car prices have been varying in short span of time. So, buyers and sellers are not able to analyse the market condition or make the proper valuation of car price. This has been resulting in over-pricing and under-pricing of the cars in the used car market.

Data Science as always, will be a great way to analyse and identify the changing trends in the domain This will be helpful in identifying the patterns for the proper valuation and making predictions of car price according to various crucial factors which are affecting the car price using the historical data and theory of probability. With the help of machine learning, we can create a model that can predict the price of car by using several inputs which consists of prominent features of the cars.

**Review of Literature**

Predicting the Price of Used Cars using Machine Learning Techniques (2014), Sameerchand Pudaruth

In this paper, 4 different machine learning algorithms have been used to predict the used cars price in Mauritius. The price attribute had to be classified into categories which contained a range of prices, but this introduced further grounds for inaccuracies. The main limitation of this study is the short of availability of data that have been used. As a further study, the author intends to collect more data and to use more advanced techniques like artificial neural networks, fuzzy logic, and genetic algorithms to predict car prices.

USED CAR PRICE PREDICTION (2021), Praful Rane, Deep Pandya, Dhawal Kotak

The focus of this project is developing machine learning models that can accurately predict the price of a used car based on its features, to make informed purchases. The author implements and evaluate various learning methods on a dataset consisting of the sale prices of different makes and models. The project has compared the performance of various machine learning algorithms like Linear Regression, Ridge Regression, Lasso Regression, Elastic Net, Decision Tree Regressor and choose the best out of it. User Interface has also been developed which acquires input from any user and displays the Price of a car according to user’s inputs.

Car Price Prediction Using Machine Learning (2021), Ketan Agrahari, Ayush Chaubey, Mamoor Khan, Manas Srivastava

The main purpose of this study is to compare the accuracy of two different prediction models for estimation of used car retail price. With the use of machine learning-based methodologies for predicting the car price based on the characteristics. To illustrate the findings, the authors have constructed a responsive website that includes all the countless used car listings. The efforts of authors culminated in this deployed service, which integrates data, machine learning, and features. This methodology can assist consumers looking to purchase a used car in making more informed judgments. Customers can now look for all automobiles in a region without physical efforts, anytime and from any location.

In this research, we used linear regression and lasso regression to develop a price model for used automobiles in comparative research. The main goal of this study is to discover the best predictive model for estimating the price of a used car.

**Motivation for the Problem Undertaken**

Automobile industry is growing every day. With the introduction of new cars to the market, and with the implementation of advanced features, the buyers are always interested in adapting to the latest technologies. But they always find it difficult to sell the existing cars. Most of these used cars are sold for under-price which is not beneficial for the user. The market-makers were using statistics and manual processing to find the patterns, trends and making the predictions in the used car price valuation based on several factors. The patterns are always changing according to various economic and market conditions. With the new features dropping every year, the demand for cars with specific features, are changing over time. But these heuristic approaches were prone to errors and changes.

With the help of machine learning and data science, which is the application of technology in statistics, the market makers and buyers can make the used car price valuation with the machine learning models. Machine learning helps Machine learning enables to analyze the valuation of car price with the help of powerful machine learning algorithms that can learn the patterns and correlations of several factors affecting the valuation of car price with the help of historical data and theory of probability.

These models always have an advantage of recalibrations with the changing trends and patterns of the used car market. Because this domain is always fluctuating and prone to changes according to various economic and other factors. With the appropriate machine learning model which scalable according to the demand and changing trends, we will be able to automate the process of predicting the price of used cars which eliminate the risk of selling the cars as over-priced or under-priced.

**Analytical Problem Framing**

**Mathematical/ Analytical Modelling of the Problem**

As per the requirement, we have to make a new machine learning model with the new data that can predict the price of used cars. Since the data is having a features and target, we can use the supervised machine learning technique as it will help as to learn the pattern with the existing data and make prediction of target (Car Price) based on various features or characteristics of the cars. Since our target variable is numerical in nature, it is best to use regression machine learning algorithms and approaches as it is built for analysing continuous data. We have used many pre-processing techniques to clean the raw data which will be helpful for the machine to analyse and learn the patterns and correlations of the features to predict the target. We have used various regression techniques such as linear regression models, regularization techniques to identify the over fitting of the models, various boosting algorithms and ensemble methods which consists of a fusion of these algorithms which makes them powerful in machine learning and making the predictions.

**Data Sources and their formats**

The project requirement consists of two phases: -

a. Data Collection Phase b. Model Building Phase.

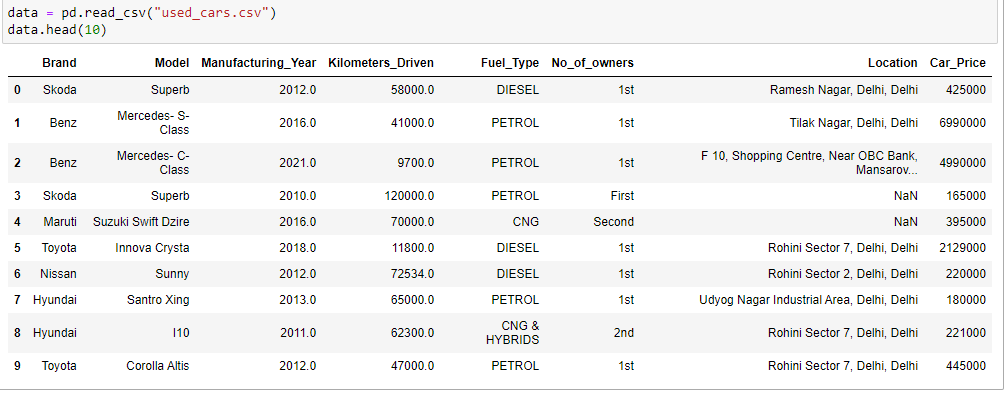
We have scraped used cars data for 5270 cars from three websites: -

a. Olx.in b. Cardekho.com c. Cars24.com

The data has included all types of cars such as SUV, Sedans, Coupe, minivan, Hatchback etc.

With the help of Selenium Web Scraping Library, we were able to scrape the required details about every car which helped us to gather the information from the three websites.

Dataset – used\_cars.csv. We have to test and train the algorithm with this dataset to create the model.



Snapshot of Dataset

* We have 5270 rows and 8 columns in the dataset.
* We have string, float, and integer type of data in the dataset.
* We have 5270 non null values in columns ['Brand', 'Car\_Price']

**Description of the dataset**

**Features in Dataset (Independent Variable)**

Brand - Brand of the car

Model - Model and variant of the car

Manufacturing\_Year - Manufacturing year of the car

Kilometers\_Driven - Total Kilometres driven as per odometer

Fuel\_Type - Fuel type

No\_of\_owners - Number of owners from the first purchase.

Location - Current location of the car or registration of the car

**Target in dataset (Dependent Variable)**

Car\_Price - Price for which the car is available for sale.

While exploring the categorical variables, we found that several categories of the columns [‘Brand’, ‘Fuel\_Type’, ‘No\_of\_owners’] contained duplicate categories which were in same names. So, we renamed the duplicate categories to show them under the original categories.

**Data Pre-processing**

**Checking for Missing Values**

Text

Description automatically generated with low confidence

Missing data in the dataset

A picture containing text

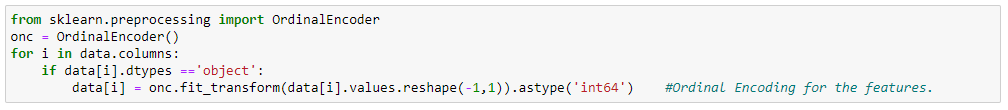
Description automatically generated

Imputation of mean and mode value for missing values in dataset

We imputed the mean value for missing values in numerical columns and mode value of column for missing values in categorical columns in dataset.

**Encoded the Categorical Columns**

Our motive is to make it easy for the machine learning algorithms to learn the patterns and correlation to make a model that can make the predictions. The machines are easy in learning numbers, so encoding the categorical variables as numeric data will make it easy for the machine to learn, train and test the data.



Encoding the Categorical Variable data

**Data Cleansing**

**Removing the skewness with power transform**



Numerical columns in dataset with skewness beyond +/-0.5 standard limit

After removing the skewness from the data using standard scaler and power transform our data looks like this:



Skewness of the dataset

**Removed Outliers in the datasets**

We have used two methods for outlier removal method:

* Outlier removal using ZScore Technique
* Outlier removal using IQR (Inter Quartile Range) Technique.

0.34% of the data loss is incurred after removing outliers using zscore outlier removal method. 0.27% of data loss is incurring after removing outliers using IQR outlier removal method. So, the IQR method is slightly better as it is losing the less data from the dataset after removing outliers.

**Checked and Removed Multicollinearity from the Datasets.**

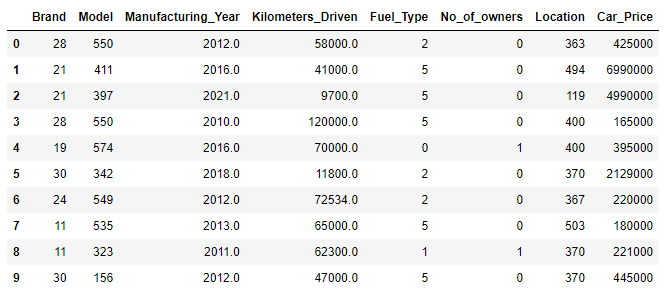
Graphical user interface, application

Description automatically generated with medium confidence

VIF Factor of Variables

We can see that none of the columns are having variance inflation factor higher than 5. That means there is no or least multicollinearity existing between the variables in the dataset.

**Final Dataset**



Snapshot of final dataset after pre-processing and data cleaning

**Data Inputs- Logic- Output Relationships**

**Statistical Summary**

**Describe of the data**

**Table

Description automatically generated**

Snapshot of describe of the data

**A screenshot of a computer

Description automatically generated with medium confidence**

Heatmap of Statistical Summary of the dataset

**Observations:**

* The mean of the columns ['Kilometers\_Driven','Car\_Price'] are higher than the median value. That means the distribution of values in these columns are not normal and skewness is present in the data distribution.
* The max value of the columns ['Kilometers\_Driven','Car\_Price'] are having enormous difference between the 75%. Possible outliers are present in the data of these columns.

**Correlation**

Chart, treemap chart

Description automatically generated

Correlation of Data Variables

None of the columns are having perfect correlation with another variable. So, we can evaluate the correlation of features with the target variable.

**Correlation with Target Variable**

Chart

Description automatically generated

Correlation of Variables with Target Variable

Manufacturing\_Year 0.251655

Model -0.007641

Location -0.024122

Brand -0.043915

No\_of\_owners -0.095251

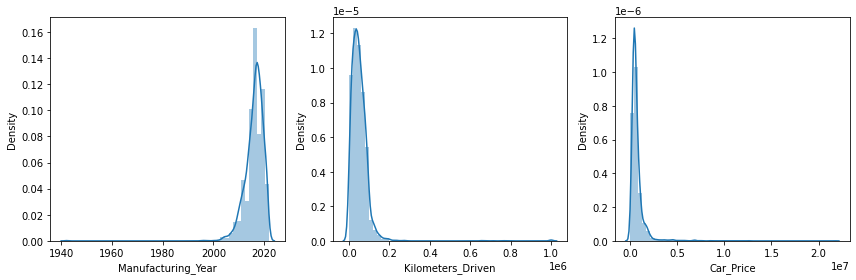
Kilometers\_Driven -0.109295

Fuel\_Type -0.173148

**Observations:**

* Except the column 'Manufacturing\_Year', rest of the columns are having a negative correlation to the target variable 'Car\_Price'.
* The column "Manufacturing\_Year' is having the highest positive correlation to the target variable 'Car\_Price'. The column 'Fuel\_Type' is having highest negative correlation to the target variable 'Car\_Price'.
* The column 'Model' is having least negative correlation to the target variable 'Car\_Price'.

**Assumptions Related to the Problem Under Consideration**



Distribution of Data in Numerical Columns

* The data in columns are not normally distributed.
* Skewness is present in the data of all numerical columns.

**Outliers**

A picture containing graphical user interface

Description automatically generated

Presence of Outliers in Numerical Variables

* Outliers are present in the data of all the numerical columns.
* Since the column 'Car\_Price' is our target variable, we will not remove the outliers from this column.

**Hardware and Software Requirements and Tools Used**

**Hardware Requirement:**

**Graphical user interface, text, application

Description automatically generated**

Hardware Configuration

**Software Requirements:**

* Windows Version: Windows 10 Pro
* Anaconda Navigator: 2.0.3

Anaconda offers the easiest way to perform Python/R data science and machine learning on a single machine.

* Jupyter Notebook: 6.3.0

Anaconda offers the easiest way to perform Python/R data science and machine learning on a single machine.

* Python3: Python 3.9.9

Python3 is used as the base environment for performing the machine learning and data analysis.

Python Libraries Used:

* Pandas: Data manipulation and analysis
* NumPy: Adding support for large, multi-dimensional arrays and matrices, along with an enormous collection of high-level mathematical functions to operate on these arrays.
* Matplotlib, Seaborn: For visualization of variable relations and data distribution, and analysis.
* Sklearn: Simple and efficient tools for predictive data analysis.
* SciPy: SciPy provides algorithms for optimization, integration, interpolation, eigenvalue problems, algebraic equations, differential equations, statistics, and many other classes of problems.
* Statsmodels: Provides classes and functions for the estimation of many different statistical models, as well as for conducting statistical tests, and statistical data exploration.
* Xgboost, catboost, lightgbm: Gradient boosting framework that uses tree-based learning algorithms.
* Pickle: Implements binary protocols for serializing and de-serializing

**Model/s Development and Evaluation**

**Identification of possible problem-solving approaches (methods)**

**Graphical user interface, text, application

Description automatically generated**

Imported the required models and created instances for the models

We created three functions for testing the model and for cross validations:

* best\_ran: Finding the best random state for the selected model
* mod\_test: Training the model with the train data using the best random state.
* cross\_val: Finding the best cross validation mean score for each model.

**Testing of Identified Approaches (Algorithms) and evaluation of selected models**

Graphical user interface, text, application, email

Description automatically generated

Code Snippet for function to find best random state

Graphical user interface, text, application

Description automatically generated

Code Snippet for function to test the model

**Graphical user interface, text, application

Description automatically generated**

Code Snippet for function to find the cross validation mean score

* Linear Regression

Table

Description automatically generated

Model Test Performance

LinearRegression ()

At cv fold 7 the cv score is 0.1602592086826202 and the R2 score is 0.21066168609496627

* KNeighborsRegressor

Table

Description automatically generated

Model Test Performance

KNeighborsRegressor ()

At cv fold 6 the cv score is 0.06286766980008612 and the R2 score is 0.4090443620841965

* SVR

Text

Description automatically generated with medium confidence

Model Test Performance

SVR ()

At cv fold 0 the cv score is 0 and the R2 score is -0.06913890396706468

* DecisionTree Regressor

Text

Description automatically generated with low confidence

Model Test Performance

DecisionTreeRegressor ()

At cv fold 0 the cv score is 0 and the R2 score is 0.7638812386427065

* RandomForest Regressor

Text

Description automatically generated

Model Test Performance

RandomForestRegressor ()

At cv fold 8 the cv score is 0.23267864095587695 and the R2 score is 0.7813948565775681

* AdaBoostRegressor

Table

Description automatically generated with medium confidence

Model Test Performance

AdaBoostRegressor ()

At cv fold 0 the cv score is 0 and the R2 score is -1.5142886069794743

* GradientBoostingRegressor

Text

Description automatically generated

Model Test Performance

GradientBoostingRegressor ()

At cv fold 9 the cv score is 0.38589722085427625 and the R2 score is 0.7843173363863506

* VotingRegressor

Text

Description automatically generated

Model Test Performance

VotingRegressor ()

At cv fold 6 the cv score is 0.4911523166557685 and the R2 score is 0.6396968402159033

* SGDRegressor

Table

Description automatically generated

Model Test Performance

SGDRegressor ()

At cv fold 7 the cv score is 0.17142521509148537 and the R2 score is 0.21093491106151396

* ExtraTreesRegressor

Text

Description automatically generated with medium confidence

Model Test Performance

ExtraTreesRegressor ()

At cv fold 8 the cv score is 0.223767983819791 and the R2 score is 0.7168263341404635

* XGBRegressor

Text, letter

Description automatically generated

Model Test Performance

XGBRegressor ()

At cv fold 8 the cv score is 0.4693459414674942 and the R2 score is 0.8497629094009553

* LGBMRegressor

Text

Description automatically generated with medium confidence

Model Test Performance

LGBMRegressor ()

At cv fold 9 the cv score is 0.4218253298532209 and the R2 score is 0.7442745416456197

* CatBoostRegressor

Text

Description automatically generated

Model Test Performance

CatBoostRegressor ()

At cv fold 6 the cv score is 0.5064668304560532 and the R2 score is 0.8336599047513329

Key Metrics for success in solving problem under consideration

* R2 Score

The proportion of the variance in the dependent variable that is predictable from the independent variable(s). It is the one of the key metrics for analysing the performance of regression models.

* Mean Squared Error

It is the average of the square of the errors. The larger the number the larger the error. **Error** means the difference between the observed value and predicted values.

* Mean Absolute Error

It is the average of difference between the measured value and “true” value.

* Regularization

Regularization is a technique used to reduce the errors by fitting the function appropriately on the given training set and avoid overfitting.

**Visualization**

**Univariate Analysis**

**Countplot – Categorical Variables**

**Chart

Description automatically generated Chart

Description automatically generated**

**Chart, histogram

Description automatically generated**

**Observations:**

* Most of the cars which are available for sale is from Brand "Maruti". Second highest cars which are available for sale is from 'Hyundai'. These are the two major car sellers in the country.
* Most of the cars which are available for sale are petrol powered. Second most cars which are available for sale are using diesel fuel type.
* Most of the cars which are available for sale are single owner vehicles.

**A picture containing histogram

Description automatically generatedHistogramplot – Numerical Variables**

**Chart, histogram

Description automatically generated**

**Chart, histogram

Description automatically generated**

**Observations**:

* Most of the cars which are available for sale are manufactured from 2008 to 2022.
* Most of the cars which are available for sale have been driven less than 1.67 lakh Kilometres.
* Most of the cars which are available for sale are having a price ranging between Rs15,000 to Rs36.3 Lakh.

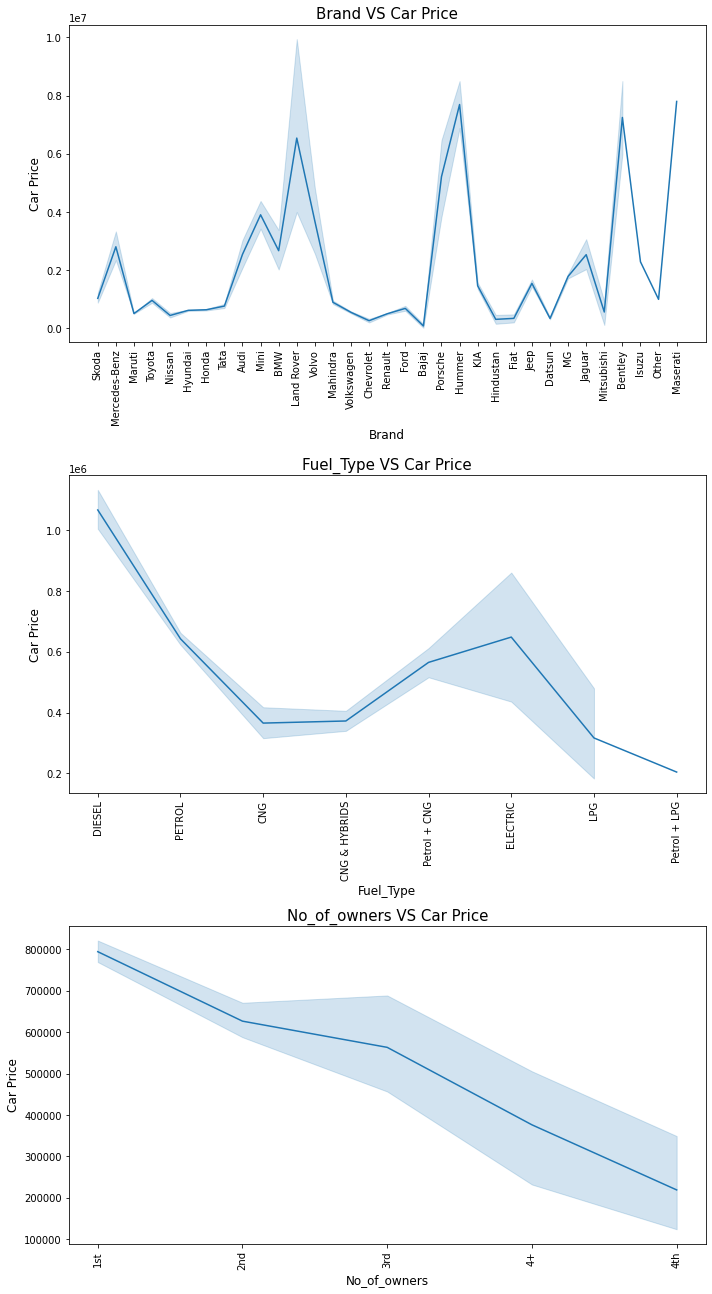
**Bivariate Analysis**

**A picture containing text, screenshot, electronics, display

Description automatically generated**

A screen shot of a computer

Description automatically generated with low confidence

****

**Chart, scatter chart

Description automatically generated**

**A picture containing shape

Description automatically generated**

**Observations**

* The sale price of car is increasing if the car was recently manufactured.
* The sale price of car is slightly decreasing for the cars which have been driven most.
* The price for the luxurious car brands like Mercedes-Benz, BMW, Land Rover, Hummer, Bentley are higher compared to other brands.
* The car price for diesel, and petrol fuelled cars are having higher price compared to other types fuelled cars.
* The price of car is decreasing as the number of owners of the cars are increasing. Most of the cars.
* Maruti is having different fuel type cars which are available for sale. Toyota and Hyundai are also having a variety of cars which are powered by several types of fuel.
* Most of the cars which are available for sale are manufactured after 2015 and most of the cars are of brands like KIA, MG which are available for sale were manufactured after the year 2019.

**Interpretation of the Results**

* Except the column 'Manufacturing\_Year', rest of the columns are having a negative correlation to the target variable 'Car\_Price'.
* The column "Manufacturing\_Year' is having the highest positive correlation to the target variable 'Car\_Price'. The column 'Fuel\_Type' is having highest negative correlation to the target variable 'Car\_Price'.
* The column 'Model' is having least negative correlation to the target variable 'Car\_Price'.

**Regularization**

**Lasso(L1)**

**Graphical user interface, text, application, email

Description automatically generated**

Lasso Regularization Performance

**Ridge(L2)**

**Graphical user interface, text, application, email

Description automatically generated**

Ridge Regularization Performance

**ElasticNet**

**Graphical user interface, text, application, email

Description automatically generated**

ElasticNet Regularization Performance

The regularization techniques didn't provide any better results.

**Hyperparameter Tuning**

**CatBoost Regressor**

**Graphical user interface, text, application

Description automatically generated**

**Text

Description automatically generated**

Best Score and parameters

Tuning the CatBoost Model

**XGBoost Regressor**

Text, application

Description automatically generated

Text

Description automatically generated

Tuning the XGBoost Model Best Score and Parameters

After all the tests, cross validations, regularization and hyperparameter tuning the XGBoost model is performing well. So, we can consider this model as the best performing model.

Graphical user interface, text, application

Description automatically generated

Snapshot of performance of the final prediction model

To test the accuracy of the model, we have made predictions on the dataset with the features of cars. Since the target variable (car price) is already available, we can cross verify the predicted values with the actual values.

Table

Description automatically generated

Chart, scatter chart

Description automatically generated

Actual and Predicted Values Line of Best Fit

Our model is performing well with predictions and provided almost accurate results. The XGBoost Regressor model(xgb) is providing a final R2 Score of 84.98%.

**Saving the best model**

**Graphical user interface, text, application

Description automatically generated**

We have saved the machine learning model for future predictions. We have serialized and saved the binary file as “Used\_car\_price\_prediction\_model.pkl” using the pickle library.

**CONCLUSION**

**Key Findings and Conclusions of the Study**

With the help of data science and machine learning, we were able to create a machine learning model using XGBoost algorithm, which can predict the used car price.

Now this model can be used to predict the price of used cars in India with the following variable information about the house. (Important Variables)

Brand  Model  Manufacturing\_Year  Kilometers\_Driven

Fuel\_Type  No\_of\_owners  Location  Car\_Price

**Impact of Variables on Target Variable (Correlation)**

* The column "Manufacturing\_Year' is having the highest positive correlation to the target variable 'Car\_Price'. The column 'Fuel\_Type' is having highest negative correlation to the target variable 'Car\_Price'.
* The column 'Model' is having least negative correlation to the target variable 'Car\_Price'.
* Except the column 'Manufacturing\_Year', rest of the columns are having a negative correlation to the target variable 'Car\_Price'. Higher the contribution of these variables for a car will result in the decrease in the car price.

**Learning Outcomes of the Study in respect of Data Science**

Used car selling market is booming as automobile industry is launching new cars with advanced tech. Since it is a normal in this era to catch up with the latest technologies, the consumers would always like to have the best and latest car. But it was always a headache when it comes to selling the old car as the market is always fluctuating and there are high chances that the car can be overpriced or under-priced.

With the help of machine learning XGBoost algorithm, we were able to create a machine learning model that can predict the price of used cars based on several features about the car. The XGBoost algorithm is based on gradient boosting which uses ensemble decision tree algorithm and since this a regression problem, the algorithm can use mean squared error as cost function. Thus, after all the tests, cross validations, regularizations, and hyper parameter tunings, we attained R2 score of 84.97%.

**Limitations of this work and Scope for Future Work**

**Limitations**

* The dataset included many key features about the used cars which helped us in the prediction of car price. But still there are many other features that has been left out which could help to improve the efficiency of the model such as milage, engine power etc.
* The used car markets are always fluctuating and always prone to changes. Even though we have included many salient features, there are other factors such as economic conditions, availability of cars service centres, spare parts prices and availability, inflation etc.
* The source of dataset was limited to three websites and popular cities. The prices may vary for other websites, and remote areas.
* The data was noisy, and we had to remove several data while the process of cleaning and pre-processing. This has altered the true nature of the data.

**Scope**

* The used car market is volatile to market and economic conditions, the trends, and patterns in this market changes over time. Since we have created a machine learning model for forecasting the used car price valuation, there is always an option for recalibration. So, we can always analyse the market and include the latest trends or features that are affecting the price of the used cars.
* Our model is scalable and can be improved by including more data and adding more features. We can also collect data from various websites and include many other unpopular cities for improving the model accuracy in prediction.
* Since the data science is growing over time, we can adapt to new features and possibilities that will help us creating a powerful model in predictions. With automation using machine learning, we are saving time, effort and incurring less errors.

**Thank You**