
# CAR PRICE PREDICTION

Submitted by:

UJJWAL PRATIK

# ACKNOWLEDGMENT

For this study, I sincerely thank faculties of Data Trained for their detailed teaching method. Topic covered like Machine Learning and Neural language with python has helped a lot. There are also some references taken from YouTube videos and google search.

# 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. We must make car price valuation model

- ## . Conceptual Background of the Domain Problem

  You have to scrape at least 5000 used cars data. You can scrape more data as well, it's up to you. more the data better the model In this section You need to scrape the data of used cars from websites (Olx, cardekho, Cars24 etc.) You need web scraping for this. You must fetch data for different locations. The number of columns for data doesn't have limit, it's up to you and your creativity. Generally, these columns are Brand, model, variant, manufacturing year, driven kilometres, fuel, number of owners, location and at last target variable Price of the car. This data is to give you a hint about important variables in used car model. You can make changes to it, you can add or you can remove some columns, it completely depends on the website from which you are fetching the data. Try to include all types of cars in your data for example- SUV, Sedans, Coupe, minivan, Hatchback.

- ## Review of Literature

  The first paper is Predicting the price of Used Car Using Machine Learning Techniques. In this paper, they investigate the application of supervised machine learning techniques to predict the price of used cars in Mauritius. The predictions are based on historical data collected from daily newspapers. Different techniques like multiple linear regression analysis, k-nearest neighbours, naïve bayes and decision trees have been used to make the predictions.

- ## Motivation for the Problem Undertaken

  Deciding whether a used car is worth the posted price when you see listings online can be difficult. Several factors, including mileage, make, model, year, etc. can influence the actual worth of a car. From the perspective of a seller, it is also a dilemma to price a used car appropriately. Based on existing data, the aim is to use machine learning algorithms to develop models for predicting used car prices.

# Analytical Problem Framing

- ## Mathematical/ Analytical Modeling of the Problem

  I have used various Mathematical and Statistical approach for making this project predict better. I have used Statistical summary for checking and analysing various stats like, number of values counts in dataset of each column, mean, median and mode for each column. Other than that, stats show outliers and skewness in overall dataset. After analysing the data through Statistical Summary, I have applied some Mathematical tools for removing outliers and skewness present in the dataset, Mathematical tools is also used for checking null values and at last making prediction.

- ## Data Sources and their formats

  I have scrapped the data of cars from various websites like CarDekho.com, Cars24, Olx etc. The Datasets contains 11000 rows and 8 columns.

  Dataset format using Pandas

| | Unnamed: 0 | Car Name and Model | Variant | Other Details | Discount Percentage | EMI | Discount | Price |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 2014 Maruti Alto 800 | VXI Manual | 12,535 km | ₹28,000 OFF | ₹6,531/month | ₹3,21,599 | ₹2,93,599 |
| 1 | 1 | 2021 Hyundai VENUE | S MT 1.2 KAPPA | 1st Owner | ₹14,000 OFF | ₹17,678/month | ₹8,08,699 | ₹7,94,699 |
| 2 | 2 | 2014 Hyundai Grand i10 | SPORTS 1.2 VTVT Manual | Petrol | ₹35,000 OFF | ₹8,386/month | ₹4,11,999 | ₹3,76,999 |
| 3 | 3 | 2017 Maruti Alto K10 | VXI Manual | 2,589 km | ₹29,000 OFF | ₹7,843/month | ₹3,81,599 | ₹3,52,599 |
| 4 | 4 | 2011 Maruti Alto K10 | VXI Manual | 1st Owner | ₹9,000 OFF | ₹4,734/month | ₹2,21,799 | ₹2,12,799 |

- ## Data Preprocessing Done
  - Importing Libraries
  - Loading the dataset
  - Checking the shape of the data
  - Checking data types
  - Checking Null values
  - Analysing through Statistical Summary
  - Univariate Analysis
  - Bivariate Analysis
  - Multivariate Analysis
  - Splitting the independent and target variable
  - Checking and removing skewness and outliers
  - Building models

- Data Inputs- Logic- Output Relationships
  - I am finding partial correlation between each input with the output, also finding high value of coefficient.
  - By extracting weights of the trained model, used connection weights method the relative importance of each input variable on the outputs can be reduced.
  - I have got the validation accuracy almost equal to the training accuracy.

- Hardware and Software Requirements and Tools Used
  - Hardware Requirement:
    - Intel core i5
    - 8 GB Ram
  - Software Requirement:
    - Python 3.x with packages
    - Pandas: Data analysis and manipulation tool
    - NumPy: Provide support for mathematical functions, random number etc
    - Selenium: Web scrapping tool
    - Chrome driver: automated Google Chrome
    - Matplotlib: is a low-level graph plotting library in python that serves as a visualization.
    - Seaborn: is a library mostly used for statistical plotting in python.
    - Scikit-Learn: is an open-source Python library that has powerful tools for data
    - analysis and data mining.

# Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)
    - R2 score: is used to evaluate the performance of a linear regression mode.
    - Linear Regression: Logistic regression is fast and relatively uncomplicated, and it is convenient for you to interpret the results.
    - Lasso: The Lasso is a linear model that estimates sparse coefficients with l1 regularization.
    - Ridge: Ridge regression is an extension of linear regression where the loss function is modified to minimize the complexity of the model.
    - Elastic Net: is a linear regression model trained with both l1 and l2 -norm regularization of the coefficients.
    - Cross-Validation-Score: a model that would just repeat the labels of the samples that it has just seen would have a perfect score but would fail to predict anything useful on yet-unseen data.
    - Grid Search CV: This function helps to loop through predefined hyperparameters and fit your estimator (model) on your training set.
    - Mean Squared Error: this metric gives an indication of how good a model fits a given dataset.
    - Root Mean Squared error: is a frequently used measure of the differences between values (sample or population values) predicted by a model or an estimator and the values observed.
    - Label Encoder: Label Encoding refers to converting the labels into numeric form.
    - Standard Scaler: Standard Scaler. Standard Scaler helps to get standardized distribution, with a zero mean and standard deviation of one (unit variance).
    - Random Forest Regressor: A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting

- Testing of Identified Approaches (Algorithms)
  - Train Test Split
  - Linear Regression
  - Lasso Regularization
  - Ridge Regularization
  - Elastic Net Regularization
  - Grid Search CV
  - Cross Validation
  - Random Forest Regressor

- Run and Evaluate selected models
  - Linear Regression

```
r2 score 1.0
error
mean absolute error 9.987001125079008e-14
mean squared error 1.6163306711531168e-26
root mean squared error 1.2713499404778832e-13
```

  - Lasso Regularization

```
r2 score 0.9999501805042103
error
mean absolute error 0.8592222696470111
mean squared error 0.9973968244184441
root mean squared error 0.9986975640395065
```

  - Ridge Regularization

```
r2 score 0.9999999801785396
error
mean absolute error 0.016560551353461093
mean squared error 0.00039682982150776255
root mean squared error 0.01992058788057628
```

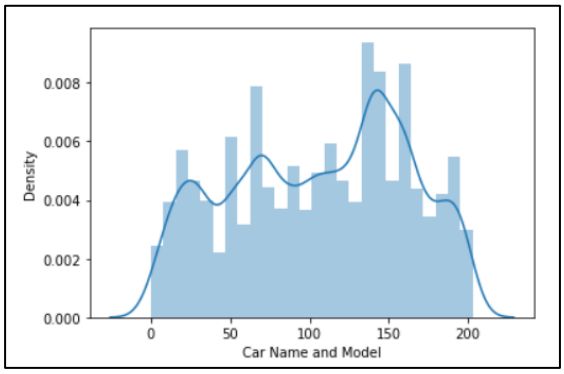
  - Elastic Net Regularization

```
r2 score 0.8950138985579602
error
mean absolute error 39.38031655101981
mean squared error 2101.843917253328
root mean squared error 45.84587132178129
```

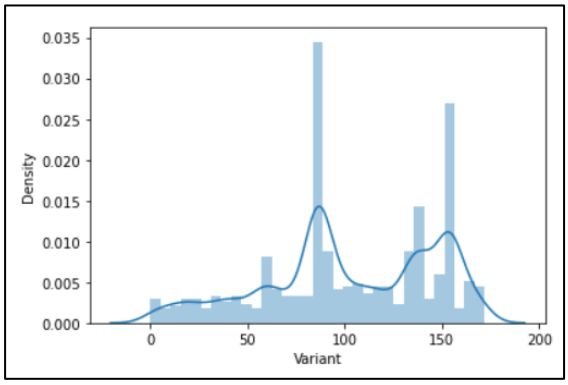
  - Random Forest Regressor

```
r2 score 99.95930931496511
cv score 99.964565936423
```

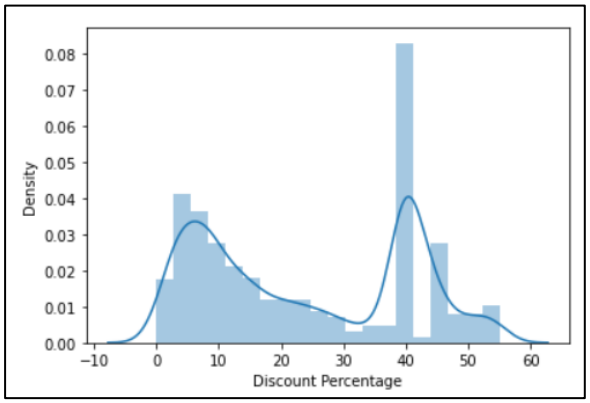
- Key Metrics for success in solving problem under consideration
  - Analysed data for any outliers
  - Analysed data for any skewness
  - Done Label Encoding for converting string value to numerical form.
  - Cross Validation for cross validates the accuracy-score from overfitting.
  - Hyper parameter tuning using Grid Search Cv for making the prediction better
  -
- Visualizations
  - Car Name and Model



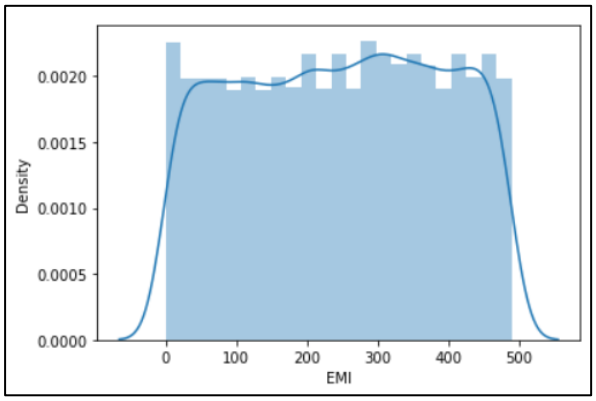

  - Variant



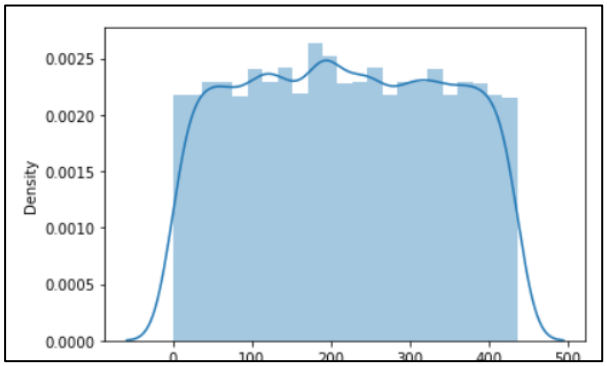

  - Discount Percentage



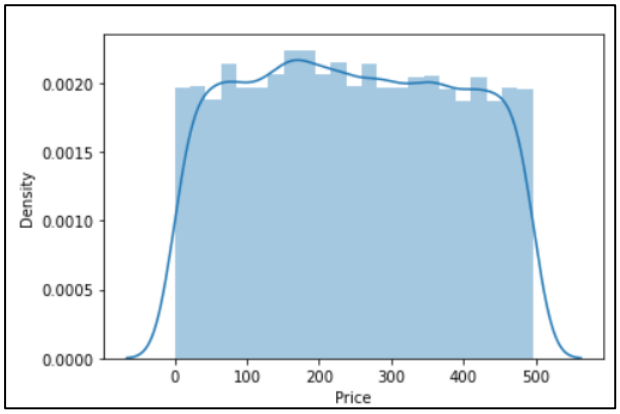
- EMI




- Discount




- Price




- ## Interpretation of the Results

  Given these considerations, we are able to predict if the price of a car with 98% of accuracy after cross validation and hyper parameter tuning. This means our model can predict the price of a cars so that buyer and seller can focus their area in very right directions to get the best out of it.

# CONCLUSION

- ## Key Findings and Conclusions of the Study

  During my analysis I got the highest accuracy with Lasso Regularization that is 99% and Ridge regularization that is also 99%, but it can be due to overfitting also. With hyper parameter tuning I got minimum difference in accuracy and cross validation is for Random Forest regressor.

- ## Learning Outcomes of the Study in respect of Data Science

  This study gives me opportunity for lots of learning starting from various types of plotting like histograms, boxplot, scatterplot, line chart and many more graphs. These graphs helped me to analyse different aspects of data like outlier, skewness, correlation etc. It also helped me to learn how to apply various model techniques on data and enable predications

- ## Limitations of this work and Scope for Future Work

  For better performance, we plan to judiciously design deep learning network structures, use adaptive learning rates and train on clusters of data rather than the whole dataset. To correct for overfitting in Random Forest, different selections of features and number of trees will be tested to check for change in performance.