Used Car Price Prediction

Aryan Raj Chhetri
CSE
PES University
Bangalore, India
aryanchhetri123@gmail.com

Kshitiz Kumar
CSE
PES University
Bangalore, India
kshitizkumar80@gmail.com

Shreyas M R
CSE
PES University
Bangalore, India
shreyasmr38@gmail.com

***Abstract*—The estimation of price of used car is one of the interesting fields of present research. Due to the steady increase of production of consumer cars with over 75 million manufactured in year 2020 which has risen to booming industry of used car market. With recent development of online portals has given rise to the need for the seller and buyer on these portals to be informed on current resale value of cars in the used car market. With of machine learning algorithms like linear regression, random forest regressor, accurate predication of resale value of used cars can be estimated.**

Keywords—* *Machine Learning, Prediction, Linear regression, Random Forest regressor.

I. Introduction

The Prediction of resale value of used car is complicated process as resale value of cars depends on various factors. The used car market has continued to increase in contrast to new car marker as sales of 2.8 million units of new cars have been recorded whereas sales of 4.4 million units of used cars have been recorded in FY2020. The used car market provides a means of business to both sellers and buyers as buyers usually resell the used car for either profit or convenience. The resale value of these car depends on various factors used model, type of transmission, type of fuel used, number of seats, type of engine, color, and another factor. The resale value of a used car is usually never a constant as the used car market always an influx of used cars every year. The resale value of cars can either depreciate or appreciate depending on the type of car.

The prediction of resale value of used cars is estimated using regression algorithms as it provides continuous values which is helpful in predicting actual resale value of cars rather than predicting a range of values for resale value. We use regression algorithms such as linear regression, random forest regressor to predict the value and choose the best of these algorithms

II. Literature Survey

Kiran S [1] This paper proposes to predict the resale value of used car by using linear regression model. The paper uses linear regression model to calculate relation between no of cylinders in car and resale value of the car. Using this approach, the model has an error rate of 10.7%. The limitation of this model is that the resale value of the car depends on many factors other than no of cylinders in the engine in the car. To make to the model to accurate a multilinear regression algorithm could be used to decrease the error rate of the model.

K. Samruddhi [2] This paper proposes to used K nearest neighbors algorithm to predict the resale value of used car. This paper, the author has selected a small dataset of used cars and the model has been trained of different ratios of training and test data set. The model has also been cross validated for assessing performance using K-fold method. This model has accuracy of 85% with Root-Mean Squared Error (RMSE) rate of 4.01 and Mean Absolute Error (MAE) rate of 2.01 with value of as 4.

III. Proposed model

The proposed model is to use various regression algorithms such as linear regression, Lasso regression, Random Forest regressor to predict the resale value of used car and to compare the accuracy and error rates of the algorithms used and select the best of the used algorithms

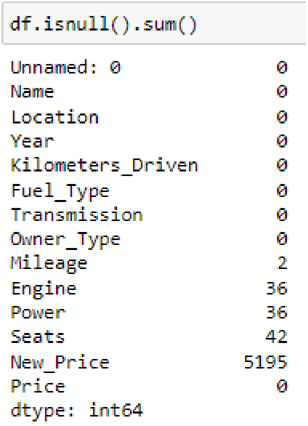
Linear Regression is used to model the relationship between two variables by fitting a linear equation to observed data. The other is dependent variable. For Example: A modeler might want to relate weights of individuals to their heights using a linear regression model. Linear regression is useful for finding relationship between multiple continuous variables

Lasso Regression. The “LASSO” stands for Least Absolute Shrinkage and Selection Operator. Lasso regression is a regularization technique. It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean. The lasso procedure encourages simple, sparse models (i.e., models with fewer parameters). This regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.

Random Forest Regressor. It is a meta estimator that fits several classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. It can be used for both classification and regression problems

IV. Data preprocessing and visualization

The data set used by the proposed model has to be cleaned before the model can be trained on it to get accurate predication. Data cleaning consists of removal of Non numerical data from numerical attributes, filling missing values with appropriate value like mean, median as missing values in engine attributes is filled with mean value of that attribute.



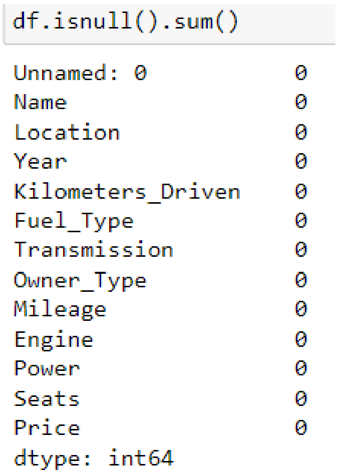
```
df.isnull().sum()

Unnamed: 0               0
Name                     0
Location                 0
Year                     0
Kilometers_Driven        0
Fuel_Type                0
Transmission             0
Owner_Type               0
Mileage                  2
Engine                  36
Power                   36
Seats                   42
New_Price             5195
Price                    0
dtype: int64
```

Fig 1

Fig shows the missing values in the attributes of dataset before data pre-processing and data cleaning.

The missing data of Engine attribute is replaced with Nan

New_price attribute is dropped because it has too many missing values, Mileage attribute is replaced with mean Mileage, Engine attribute is replaced with mean data, Seat attribute is replaced with mean data



```
df.isnull().sum()

Unnamed: 0           0
Name                 0
Location             0
Year                 0
Kilometers_Driven    0
Fuel_Type            0
Transmission         0
Owner_Type           0
Mileage              0
Engine               0
Power                0
Seats                0
Price                0
dtype: int64
```

Fig 2

Fig 2 shows the missing values in the attributes of dataset before data pre-processing and data cleaning.

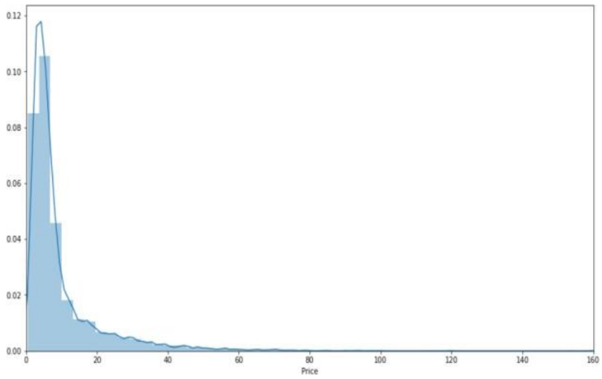


Fig 3: Graph of Price of Used cars

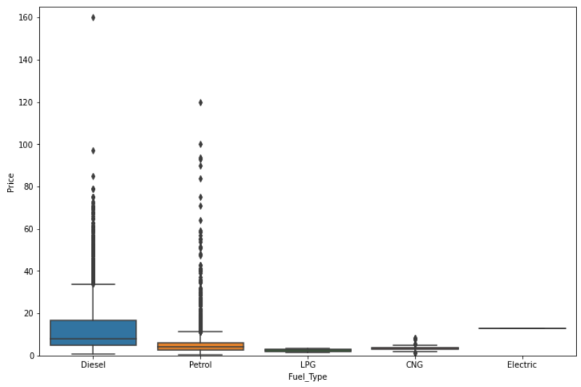


Fig 4: Boxplot of Fuel types vs Price

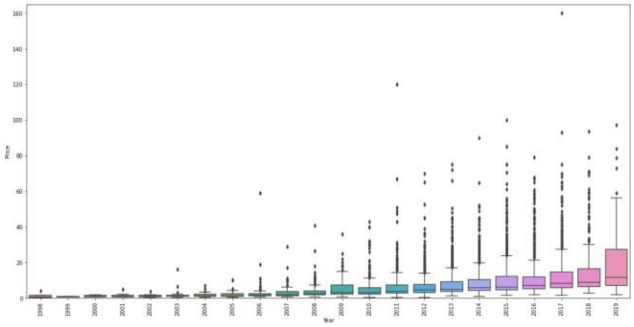


Fig 5: Boxplot of Year vs Price

Fig 5 shows most of the cars are manufactured in the last 5 years are more expensive than older cars

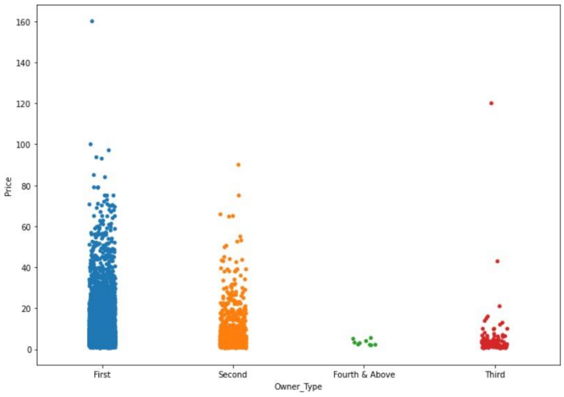


Fig 6: Strip plot of Owner type vs Price

Fig 6 shows that most of the cars are firsthand cars

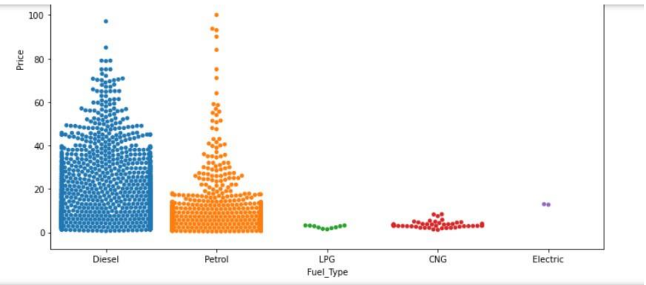


Fig 7: Swarm plot of Fuel type vs Price

Fig 7 shows that most of the car’s fuel type is diesel which are being resold

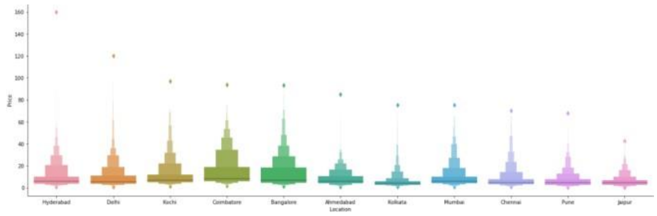


Fig 8: Boxplot of Location vs Price

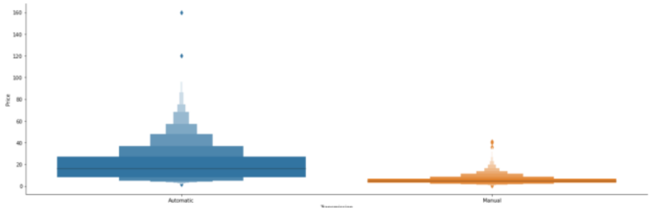


Fig 9: Plot of Transmission vs Price

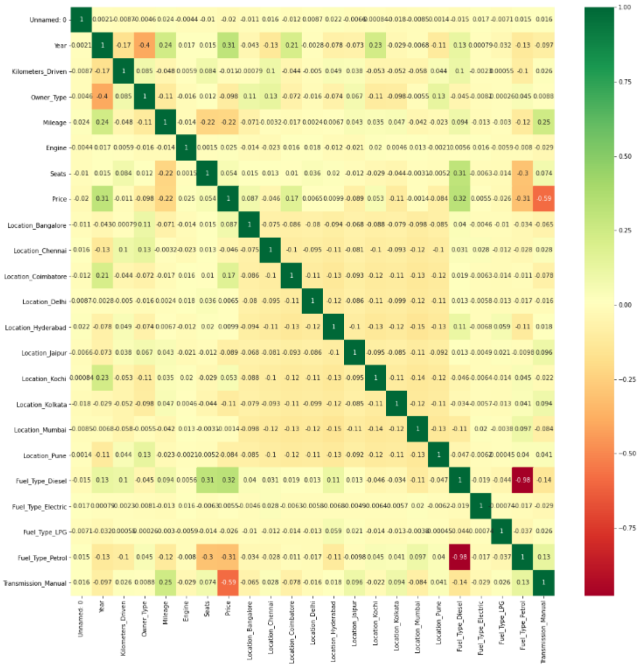
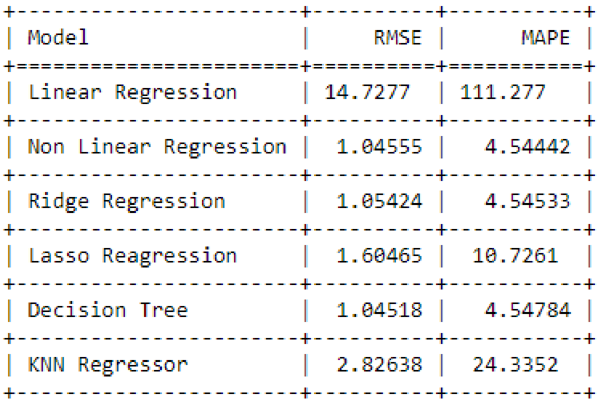


Fig 10: Heatmap of Attributes

Fig 10 shows the correlation coefficients of numerical variable.

V. PROPOSED SOLUTION



```
+-----------------------+----------+-----------+
| Model                 |     RMSE |      MAPE |
+=======================+==========+===========+
| Linear Regression     | 14.7277  | 111.277   |
+-----------------------+----------+-----------+
| Non Linear Regression |  1.04555 |   4.54442 |
+-----------------------+----------+-----------+
| Ridge Regression      |  1.05424 |   4.54533 |
+-----------------------+----------+-----------+
| Lasso Reagression     |  1.60465 |  10.7261  |
+-----------------------+----------+-----------+
| Decision Tree         |  1.04518 |   4.54784 |
+-----------------------+----------+-----------+
| KNN Regressor         |  2.82638 |  24.3352  |
+-----------------------+----------+-----------+
```

Fig 11: RMSE and MAPE comparison of all Regression models

Table in Fig 11 shows Decision Tree has lowest root mean square error among all the regression models. This means Decision Tree fits the data set better than the other models.

From the table we observe that Random Forest Regressor model has the lowest mean absolute percentage error. This means this models is more accurate than the other regression models.

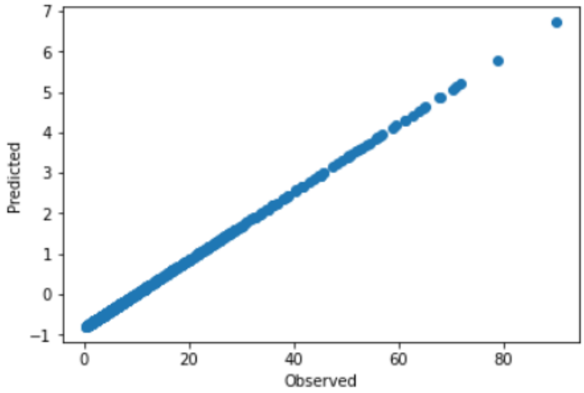


Fig 12: Observed vs Predicted Plot of Linear Regression

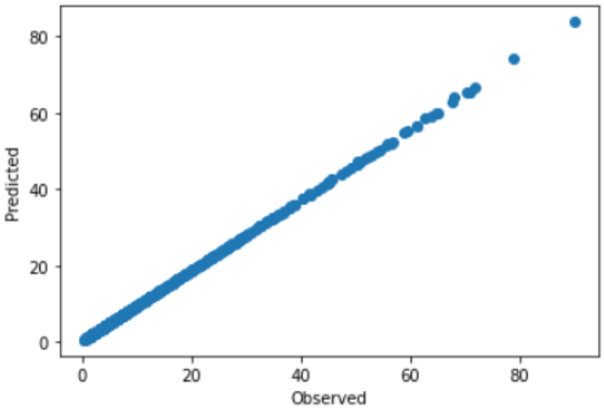


Fig 13: Observed vs Predicted Plot of Random Forest Regression

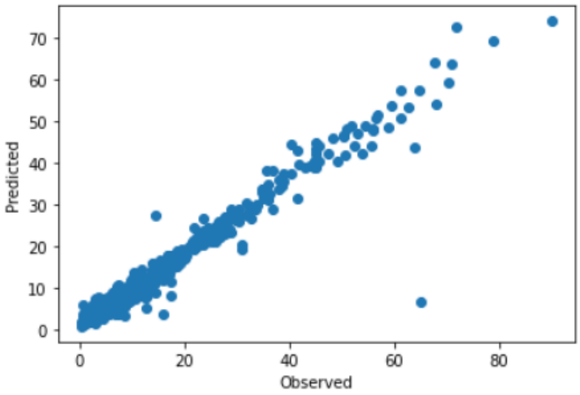


Fig 14: Observed vs Predicted Plot of Ridge Regression

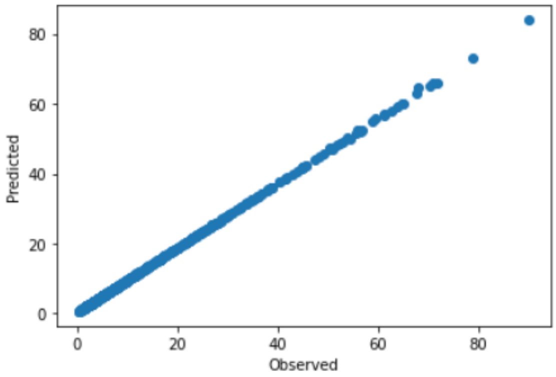


Fig 16: Observed vs Predicted Plot of Decision Tree

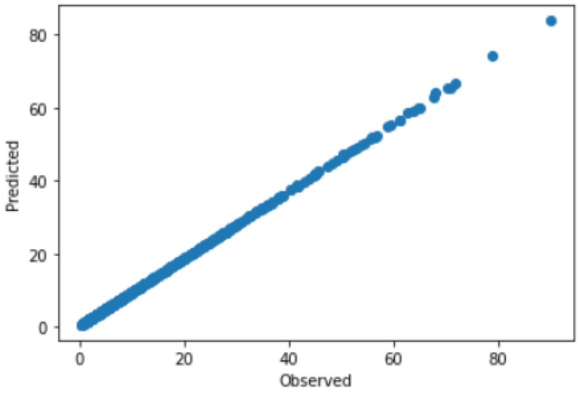


Fig 15: Observed vs Predicted Plot of Lasso Regression

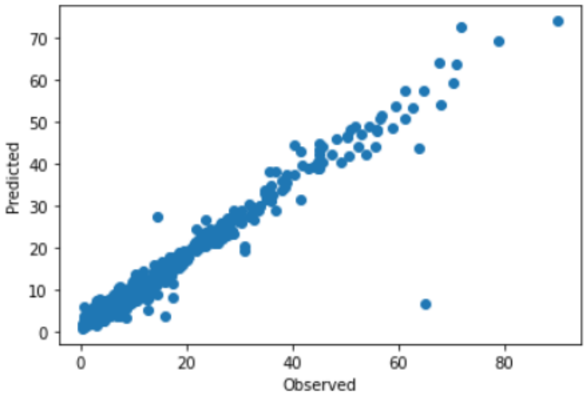


Fig 17: Observed vs Predicted Plot of KNN Regression

Data set was split into testing and training set in 1:4 ratio and all six regression models were applied to it to find the RMSE and MAPE value.

Linear Regressor:

Linear Regression model was used to predit the car price of used cars by importing LinearRegression from sklearn using python. MAPE value was found to be 111.2774 and RMSE value was found to be 14.7277.

Random Forest Regressor:

This model was used by importing RandomForestRegressor from skleran.ensemble in python. In this supervised learning algorithm ensemble method was used to solve both regression and classification problem.MAPE and RMSE values were found to be 4.5444 and 1.0456 respectively.

Ridge Regressor:

In this regression model multicollinearity issue is handled by importing Ridge from sklearn.linear _model. RMSE in this model is 1.0542 and MAPE is 4.5453.

Lasso Regression:

Even this model is used in case of multicollinearity. It provides regression coefficients which can be regularized to avoid overfitting. This model was implemented by importing liner_model. Lasso from sklearn in python. RMSE of this model is 1.6047 and MAPE value is 10.7261

Decision Tree:

This regression method builds classification models in the form of a tree structure by breaking dataset into smaller subsets. This model was used for predicting used car price by importing DecisionTreeRegressor from sklearn.tree.RMSE and MAPE values are found to be 1.0452 and 4.5478 respectively.

KNN Regressor:

In this method the target is predicted by local interpolation of the targets associated of the K nearest neighbours in the training set.This ML model is implemented by importing KNeighboursRegressor from sklearn.neighbors in python.

Best K value was found for this data set which turned out to be 5.RSME and MAPE for K=5 are 2.8264 and 24.3352.

Conclusion

The dataset has been pre-processed and cleaned and is suitable for the model to train on this dataset. From data visualization we know the relation between price and other attributes of a used car. Six ML Regression models were used to predict the used car price of different models. We have found Decision Tree regressor to be with lowest RSME value. Random Forest Regressor has lowest mean absolute percentage error. This concludes that Decision Tree regressor and Random Forest Regressor models are more accurate than other ML regression models used for this particular data set.

Reference

[1] Kiran S," Prediction of resale value of car using linear regression algorithm"," International Journal of Innovative Science and Research Technology Year:2020 | Conference Paper | Publisher: IJIST"

[2] K. Samruddhi," Used Car Price Prediction using K-Nearest Neighbor Based Model"," International Journal of Innovative Research in Applied Sciences and Engineering (IJIRASE) Year: 2020 | Conference Paper | Publisher: IJIRASE"