# Car Price Prediction Using Ensemble Method

Vankayala Sri Sailaja
Dept.of Computer Science and Engineering
Sharda University
Greater Noida,India
2019008425.vankayala@ug.sharda.ac.in

Aman Kumar Kasaudhan
Dept.of Computer Science and Engineering
Sharda University
Greater Noida,India
2019008353.aman@ug.sharda.ac.in

Nibbritta Niloy Sarker Tanmoy
Dept.of Computer Science and Engineering
Sharda University
Greater Noida,India
2019008292.nibbritta@ug.sharda.ac.in

***Abstract*—Used car price forecasting is a daunting task due to its complex non-linear and chaotic behavior.During the past few decades,both sellers and buyers devoted proactive knowledge to solving this issue.Some of them are focused on key factors that can affect car price prediction accuracy This paper expands on this particular branch of recent work by considering several influential features as inputs for testing car price forecasting performance in the period of COVID-19.The empirical results indicate that the proposed methods are efficient and warrant further research in this particular area.**

***Keywords— Used Car price prediction; Regression Techniques;XG Boost***

Despite tremendous developments in the auto industry,there are many people who prefer used cars,feeling that they are of higher quality,and are less expensive,moreover they may prefer older cars because of their ease of handling.They are not having sophisticated technology as they do,which makes it easier for them to use.According to the survey,[1]the used car market in India was worth US$27 billion in 2020,and is expected to reach US$50 billion by 2020.The COVID-19 pandemic had negligible impact on the industry.The decrease in the rate of cash flow due to the pandemic has forced buyers to opt for second hand cars,With the pandemic hampering sales and production of new vehicles,the only option for buyers is the used cars.The pre-owned car market recorded sales of 4.4 million units in FY20,while new commuters sold only 2.8 million units.vehicle in the same year.So these facts look at the importance of the developed used car market in the country.

The researchers provided some possible evidence for this effect.They checked that the price usually depends on various features which is the most dominant brand and model,age,kms driven and mileage.[2]The odometer reading also has a significant impact on a car's valuation.A car that is old on paper has become obsolete as per the norms of the country's automotive authorities,which would make it unfit to drive on the road,be it the odometer reading.Hence,a car that has been used less and has only 2-3 years of life left is considered as scrap,until we come up with measures that will practically last till the condition of each car.and provide fitness certificate accordingly.So in this paper we have applied various methods and techniques to get high accuracy of used car price prediction.

The main objective of our research is to apply a machine-learning approach to forecasting used car prices and to address the following questions:[3]Can machine learning based models make accurate predictions of a used car?Which features selection can highly impact performance?Can we get comparable or relatively high prediction performance by introducing the ensemble method?The structure of this paper is organized as follows.Section 2 depicts factors based on car price after literature review and Section 3 depicts the role of feature selection,prior to that how data is prepared and the various regression models are applied and their outcomes.Section 4 contains results and hence prooving that XGBoost(Ensemble Method)giving best MSLE(Mean Square Log Error)and RMSLE(Root Mean Square Log Error)Section 5 breifs about the process from data preprocessing to experimental

## II.Literature Review

The Review of the researchers working on this topic are as follows:Ela kumar and Piyush Kumar Yadav discussed that[4]they used automatic machine learning techniques.They used Object Detection using various categorical features.Their outcome was that Linear and Random Forest Regression has better accuracy.

ANOTHER SURVEY IS DONE BY PUDARUTH[5],HE PREDICTED THE PRICE OF USED CARS IN MAURITIUS USING MULTIPLE LINEAR REGRESSION,K-NEAREST NEIGHBORS,NAIVE BAYES AND DECISION TREES.THOUGH THEIR RESULTS WAS NOT SUFFICIENT FOR EVALUATION OF PRICE DUE TO USING LESS NUMBER OF OBSERVATIONS.PUDARUTH DISCUSSED IN HIS PAPER THAT THE DECISION TREE AND NAIVE BAYES ARE UNABLE TO USE FOR VARIABLE WITH A CONTINUOUS VALUE.

NOOR AND JAN[6]USED MULTIPLE LINEAR REGRESSION TO PREDICT VEHICLE CAR PRICE.THEY PERFORMED VARIABLE SELECTION TECHNIQUE TO FIND THE MOST INFLUENCING VARIABLES THEN ELIMINATE THE REST.IN THEIR DATA THEY CONSIDERED ONLY SELECTED VARIABLE WHICH WERE USED TO PERFORM LINEAR REGRESSION MODEL.THE RESULT WAS IMPRESSIVE WITH R-SQUARE=98%.

PEERUN ET AL.[7]DID A RESEARCH TO EVALUATE THE PERFORMANCE OF THE NEURAL NETWORK IN USED CAR PRICE PREDICTION.THE PREDICTED VALUE,HOWEVER,ARE NOT VERY CLOSE TO THE ACTUAL PRICE,ESPECIALLY ON CARS WITH A HIGHER PRICE.THEY CONCLUDED THAT SUPPORT VECTOR MACHINE REGRESSION SLIGHTLY OUTPERFORM NEURAL NETWORK AND LINEAR REGRESSION IN PREDICTING USED CAR PRICE

NITIS MONBURINON[8]CONDUCTED USING MULTIPLE LINEAR REGRESSION AND FNDED MSE OF BOTH AND CONCLUDED GRADIENT BOOSTED REGRESSION.AND MULTIPLE LINEAR REGRESSION TO GIVE BEST ACCURACY.THUS WE REALIZED THAT NONE OF THEM HAVE IMPLEMENTED USING ENSEMBLE METHOD

## III.METHODOLOGY

1.The DataSet:

The dataset used in this project has been taken fromKagglepresented in csv format.

Data Preparation:

Before processing data we have to remove unnecessary features like url,transmission,colour,county,state,owner details from the dataset.

And in next step,we will check the missing values for each feature.

Data preprocessing:

Label Encoder:In our dataset,12 features are categorical and numerical variables(excluding the value column).To implement ML model,we need to convert these categorical variables into numerical variables and LabelEncoder Library of SK Learn is used.

Normalization:The dataset is not normally distributed.All facilities have different categories.Without normalization,the ML model will attempt to disregard the coefficients of features that have low valuesbecause their effect will be much smaller than that of the larger value.So to generalize,the sklearn library i.e.MinMaxScaler is used.

Train the data:In this process,90%of the data was splitted for the train data and 10%of the data was taken as the test data.

4.ML Models:In this section,various machine learning algorithms are used to predict the value/target-variables.

The datasets are monitored,so the models are applied in the following order:

linear regression

ridge regression

lasso regression

k-neighbor registrar

random forest registrar

bagging registrar

adaboost registrar

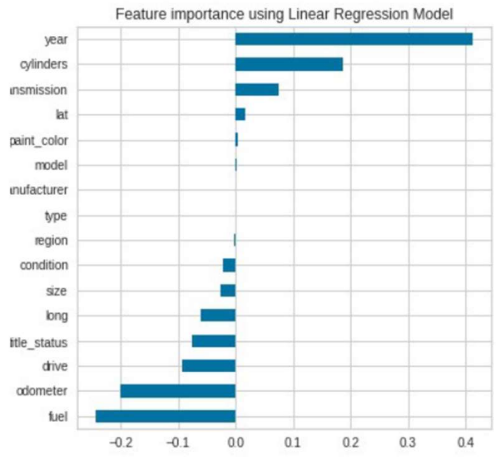
XGboost

1) Linear Regression:

A linear technique to modelling the relationship between a scalar response (or dependent variable) and one or more explanatory variables is known as linear regression in statistics (or independent variables). Relationships are modelled using linear predictor functions whose unknown model parameters are derived from data in linear regression. Linear models are a type of model that fits this description

Coefficients: The direction of the link between a predictor variable and a responder variable is shownby the sign of each coefficient.

There are two signs Positive and Negative indicating directly proportionality of the predictor and response variable.




Graph showing important feature of the dataset

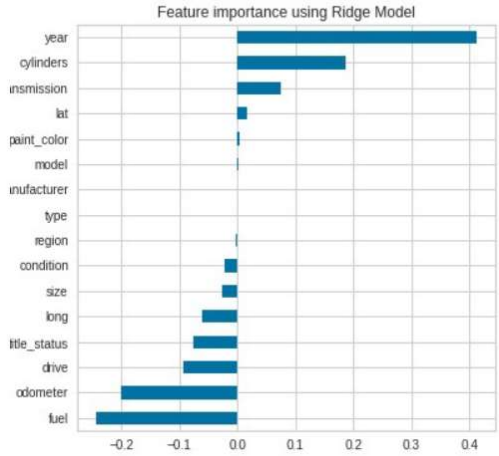
Considering this figure,linear regression shows that the five variables are the year,cylinder,transmission,fuel,and odometer.

**Result of Linear regression:**

| | |
|---|---|
| Mean Squared Log Error | 0.0024339992647452137 |
| Root Mean Squared Log Error | 0.04933557808260904 |
| R2 Score | 0.5930 or 59.30% |

2)Ridge Regression:

In cases where the independent variables are highly correlated, ridge regression is a method of calculating the coefficients of multiple-regression models. Yellowbrick library alpha selection was used to identify the best alpha value in ridge regression. Only dataset can be fit if the alpha value is 20.336. It is a variable, even though it is not a constant value. The Ridge regressor is implemented with this alpha value.




**Result of Ridge regression:**

| | |
|---|---|
| Mean Squared Log Error | 0.0024339951220568403 |
| Root Mean Squared Log Error | 0.04933553609779507 |
| R2 Score | 0.5930 or 59.30% |

The performance of Ridge Regression is almost the same as that of Linear Regression.

3. )Lasso Regression:

Lasso regression is a shrinkage-based linear regression. When data values shrink towards a central point, such as the mean, this is known as shrinkage. The Lasso method favours the use of simple, sparse models (i.e. models with fewer parameters)

The fundamental goal of Lasso regression is to find a group of predictors that decreases the quantitative response variable's prediction error. The Lasso model accomplishes this by imposing a parameter constraint that leads the regression coefficients for some variables to decrease toward zero.

**Result of Lasso regression:**

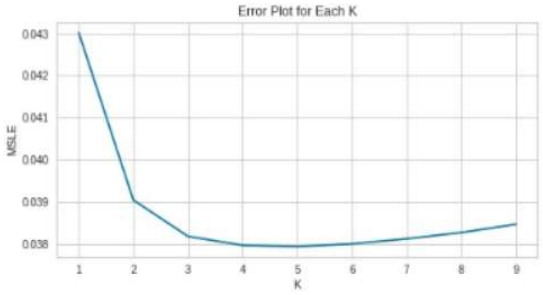
| | |
|---|---|
| Mean Squared Log Error | 0.002434007918610632 |
| Root Mean Squared Log Error | 0.04933566578663586 |
| R2 Score | 0.5930 or 59.30% |

But for this dataset, there is no need for Lasso regression as there is not much difference in error.

4) K Neighbors Regressor:

Regression-based on k-nearest neighbors. Target prediction is done by local insertion
of targets connected to the nearest neighbors of the training set.
k-NN is a type of example-based learning, or lazy learning, where the function is only locally

approximated and all computation is postponed until function evaluation.




```
K = 1  , Root MSLE = 0.043028095161555396
K = 2  , Root MSLE = 0.03904982204340027
K = 3  , Root MSLE = 0.038189303148938446
K = 4  , Root MSLE = 0.03797727462083389
K = 5  , Root MSLE = 0.03794796985755059
K = 6  , Root MSLE = 0.038011328106944124
K = 7  , Root MSLE = 0.03813151901019493
K = 8  , Root MSLE = 0.038279778248659885
K = 9  , Root MSLE = 0.03847623714879952
```

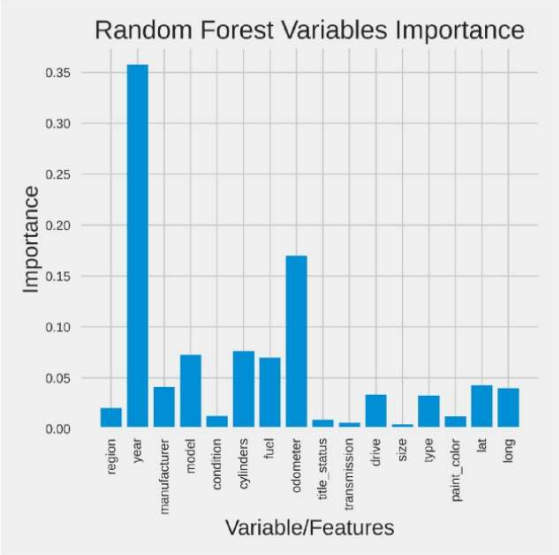
**Result of KNN:**

| | |
|---|---|
| Mean Squared Log Error | 0.0014400484163095684 |
| Root Mean Squared Log Error | 0.03794796985755059 |
| R2 Score | 76.4681% |

The performance of KNN is better and the error is decreasing with the increased accuracy.

5) Random Forest:

A random forest is a decision tree-based categorization technique. When constructing each tree, it employs bagging and feature randomization in order to produce a fragmented forest of trees whose committee forecasts are more accurate than any one tree's.
There were 180 max features 0.5 choices made in our model.




This is a simple bar plot showing that the year is the most important characteristic of a car and then the odometer variable and then the others.

.

**Result of Random Forest:**

| | |
|---|---|
| Mean Squared Log Error | 0.0007781140491380673 |
| Root Mean Squared Log Error | 0.0007781140491380673 |
| R2 Score | 0.8759 or 87.59% |

Random forest has better performance and increases accuracy to approx. 10% which is good. Since the random forest is using bagging while building each tree, the next bagging register will be done

6)Bagging Regressor:

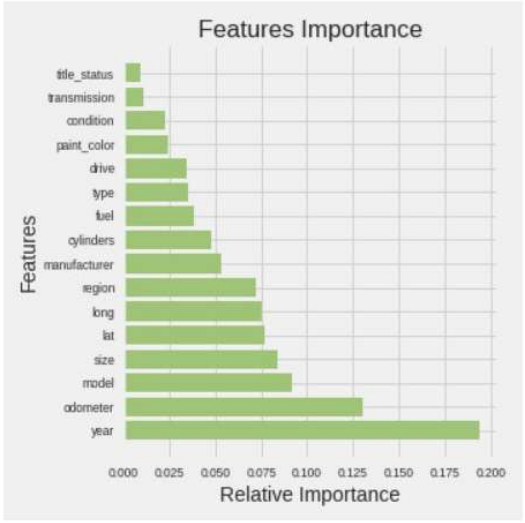
The term "bagging" refers to a form of ensemble regressor. It's a meta-estimator that fits each of the base regressors to a random subset of the original dataset and then averages or polls their predictions to create the final prediction. It is used to minimise the variance of a black-box estimator (such as a decision tree) by including randomization into the construction process and then grouping the results. DecisionTreeRegressor is employed as an estimator in this model, with a depth of 20 and 50 decision trees.

**Result of Bagging Regressor:**

| | |
|---|---|
| Mean Squared Log Error | 0.001431926503300646 |
| Root Squared Mean Log Error | 0.037840804739072954 |
| R2 Score | 0.76809 or 76.809% |

7)Adaboost regressor:

It's used to help any machine learning algorithm perform better. Adaboost assists you in combining several "weak classifiers" into a single "strong classifier."
It's a basic bar graph that indicates that the most essential feature of an automobile is the year, followed by the odometer variable, model, and so on. The DecisionTreeRegressor is utilised as an estimator in our model, with a maximum depth of 24 and 200 trees produced, and a learning rate of 0.6, as shown below.

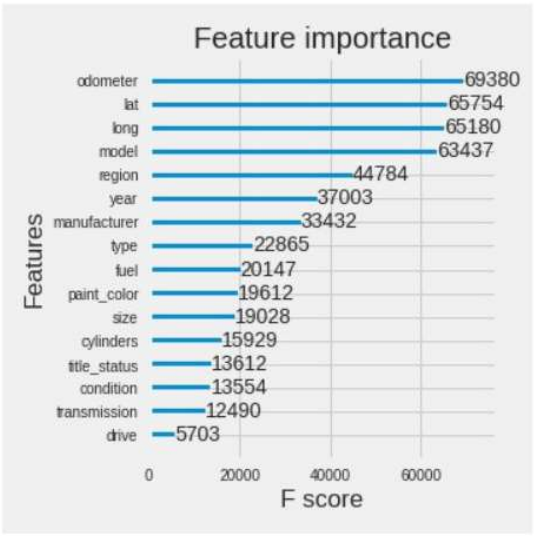



**Result of Adaboost Regressor:**

| | |
|---|---|
| Mean Squared Log Error | 0.000844759762867486 |
| Root Squared Mean Log Error | 0.029064751209454485 |
| R2 Score | 0.864084 or 86.4084% |

8)XGBoost:XGBoost stands for eXtreme Gradient Boosting

XGBoost is a method for ensemble learning. Gradient Boosted Decision Trees include XGBoost, which is utilised for speed and performance. The beauty of this strong algorithm is its scalability, which allows for quick learning via parallel and distributed computing while still conserving memory.




It's simple bar plot in descending importance which shows which attribute/variable is an important characteristic of a car which is more important. According to XGBoost, the odometer is an important feature while the year is an important feature from the previous model.

.

**Result of XGBoost Regressor:**

| | |
|---|---|
| Mean Squared Log Error | 0.0006504702126268066 |
| Root Mean Log Error | 0.02550431752913233 |
| R2 Score | 0.896623 or 89.6623% |

Car Price Prediction Using Ensemble Method


| Model | MSLE | RMSLE | Accuracy |
|---|---|---|---|
| Linear regression | 0.00243399 | 0.04933557 | 59.3051% |
| Ridge regression: | 0.00243399 | 0.04933553 | 59.3051% |
| Lasso regression | 0.00243400 | 0.04933566 | 59.305% |
| KNN | 0.00144004 | 0.03794796 | 76.4681% |
| Random Forest | 0.00077811 | 0.00077811 | 87.5979% |
| Bagging Regressor | 0.00143192 | 0.03784080 | 76.809% |
| Adaboost Regressor | 0.00084475 | 0.02906475 | 86.4084% |
| XGBoost Regressor | 0.00065047 | 0.02550431 | 89.6623% |

From the above data,we can conclude that XGBoost register is performing better than other models with 89.662%accuracy

## VI.RESULTS

Hence after performing tests using all supervised learning regression algorithms,we conclude few observations.We analyze that the diesel variant car costs more then the electric variant costs.Hybrid variant cars cost the lowest.We analyze that the car price of the respective fuel also depends on the condition of the car.Condition of the car also plays vital role in price prediction.

## V.CONCLUSION

By performing different ML models,we aim to achieve better results or less error with maximum accuracy.Our objective was to estimate the price of used cars with 25 predictions and 509577 data entries.Data cleaning is done to remove null values.Then various models were implemented,but the maximum accuracy was recieved in XG Boost(Ensemble Method).As a regression model XGBoost gave the best MSLE and RMSLE values.

## References

[1]https://www.mordorintelligence.com/industry-reports/india-used-car-market

[2]https://autoportal.com/articles/factors-which-affect-used-car-valuation-price-6446.html

[3]file:///C:/Users/sharda/Downloads/oil%20price%20prediction%20using%20ensemble%20method.pdf

[4]http://www.lingcure.org/index.php/journal/article/view/1660

[5]S.Pudaruth,“Predicting the Price of Used Cars using Machine Learning Techniques,”International Journal of Information&Computation Technology,vol.4,no.7,pp.753–764,2014.

[6]N.Kanwal“Vehicle Price Prediction System using Machine Learning Techniques”

[7]S.Peerun,N.H.Chummun,and S.Pudaruth,“Predicting the Price of Second-hand Cars using Artificial Neural Networks,”The Second International Conference on Data Mining,Internet Computing,and Big Data,no.August,pp.17–21,2015.

[8]https://sci-hub.zidianzhan.net/

[9]https://towardsdatascience.com/used-car-price-prediction-using-machine-learning-e3be02d977b2