International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 8 Aug 2021, pp: 1135-1145 www.ijaem.net ISSN: 2395-5252

Used Car Price Prediction Using Cat Boostgradientmachine and Light Gradient Boosting Machine

[1]Sai Guptha Grandhi, [2]Bangaru Suresh, [3]Chittajallu Nagarjuna,
Junior Data Scientist, Social Tek Al & ML Solutions, Hyderabad
Junior Data Scientist, Social Tek Al & ML Solutions, Hyderabad
Junior Data Scientist, Social Tek Al & ML Solutions,Hyderabad

Submitted: 01-08-2021 Revised: 14-08-2021 Accepted: 17-08-2021

ABSTRACT: On going Recession in the world, the automobile is one of the industries that suffers a lot in the production of cars because the availability of technology, designing and manufacturing of the parts are not available in the market for best price. So, people in the India shows inclined towards used cars which have innumerable benefits. A used car prediction has been a high meticulous researcharea, as it requires notable efforts and knowledge of the fieldexpert.Significantnumberofvariousattributes are examined for the reliable and accurate prediction.Tobuildamodelforpredictingtheworth of used cars in India. We applied seven machine learning techniques (Multiple linear Regression, Random Forest, Gradient Boost Machine, light GBM, Cat Boost regression, Ada Boost Regression and XGB). However, the mentioned algorithms wereappliedtoworkasanindividual,ensembleand stalking. The data used for the prediction was collected from the web portal Car24.com using web scraper that was written in Python programming language with Selenium. Performance of different algorithms were then compared to find one thatbest suits the scrapped data set. The final prediction model was integrated into Python application. Furthermore, the model was evaluated using test data and the comparing of different evaluation metricslikecoefficientofdetermination(r2)androot mean square error (RME) the lowest MAE and R2 obtained with Extreme Gradient Boosting train r2 with0.986andtestr2with0.944.IntermsofRMSE the train values are with 0.118 and test0.120.

KEYWORDS: used car price prediction, regression, machine learning, linear regression, cat boost regression, random forest, GBDT, light GBM and GBM.

I. INTRODUCTION:

The field of Artificial Intelligence has allowed analysts to cover insights from historical data and past events to get the inference and to predict the future insights. One the most infamous case study used car price prediction. Over, rapid growing populationinIndia,transportationplaysaprominent role in which car is one kind of transportation, according to study conducted by Indian Blue Book, inFY20theusedcarmarketinIndiastandpointat 4.2 million units. It also claimed that used car sales increased by 5% YoY within the in the same fiscal, while the new car sales contracted by 17.8%, due to the pre-pandemic depression. The used car sales were actually 50% above the amount of latest car sales recorded at 2.8 millionunits.

Accurate car price prediction involves expert knowledge, because price usually depends on various attributes and factors. Typically, most significant ones are brand and model, age, Kilometers travelled, Insurance, Type of Transmission, fuel type used and ration of the car. The dataset collected using web scrapping with seleniumfromcar24.com,thepurposeisaboutprice prediction of the car. We split the historical data set thatweobtainedfromcar24.comintothreegroups– trainset,validationsetandtestset.Thetraindataset are going to be input for the training model. The validation data set will be an to check model for the generations of predictions. This last data set will be used to check the correctness for the generations of predictions for the test data. There can be many flawsinthetrainsetbecauseofwhichtrainingofthe modelonthebasisofsuchdataisnotpossibleorcan leads to an incorrectoutput.

This dataset has been studied and analyzed using various machine learning algorithms like Multiple linear regression. various languages and

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 8 Aug 2021, pp: 1135-1145 www.ijaem.net ISSN: 2395-5252

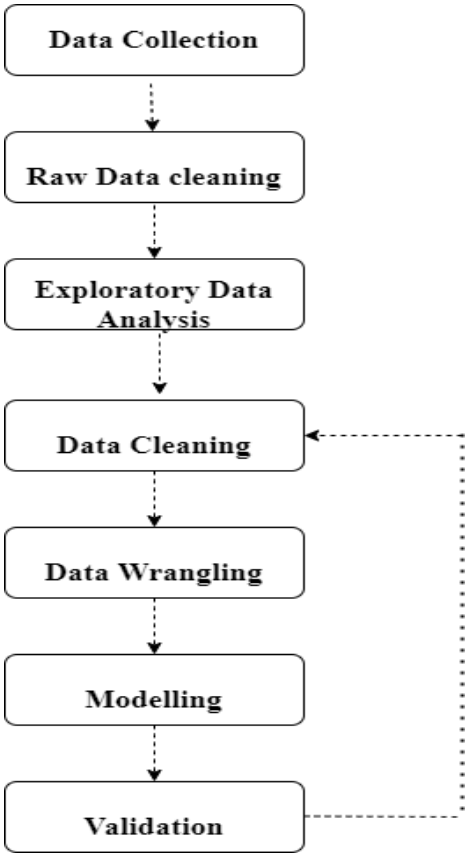
tools are used to implement these algorithms including PythonTheapproachoftheresearchpaperiscentered on Python for executing algorithms-, Random Forest, GBDT, light GBM, Non-Linear regression and XGB.

II. RELATED WORK:

Predicting price of a used cars has been studied extensively in various researches. Listian discussed, in her paper about regression model that was built using Support Vector Machines (SVM) can predict the worth of a car that has been leased with better precision than multivariate regression or some simple multiple correlation. This is often on the grounds that Support Vector Machine (SVM) is bestin handling datasets with more dimensions and it's less susceptible to overfitting and underfitting. The weakness of this research is that a change of straightforward regression with more advanced SVM regression wasn't shown in basic indicators like mean, variance orvariance.

III. EXPERIMENTAL METHODS:

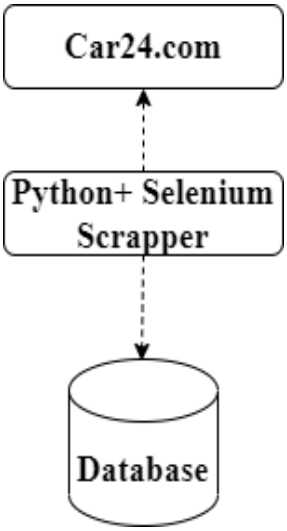
Approach for used car price prediction proposed in this paper is composed of several steps,



IV. DATA COLLECTION:

Data is collected from a local web portal for selling and buying cars car24.com, during post Pandemic, astimeintervalitselfhashighimpactonthepriceof the cars in India. The following attributes were captured for each car: brand, model, years of car purchased, kilometers travelled, fuel, year of manufacturing,transmissiontype,city,sincemanual data collection is hectic and time-consuming task, especially when there are numerous records to process, a "web scraper" as a part of this research is createdtogetthisjobdoneautomaticallyandreduce the time for datagathering.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 8 Aug 2021, pp: 1135-1145 www.ijaem.net ISSN: 2395-5252



Web scraping is the process of using automatic bots to extract content and data from a website and save data into required database. Manual data extraction is time consuming and therefore web scrapers are used to do this job in a small amount of time.

RAW DATA CLEANING:

	Name	Price	Rating	city	Kilometers	Year_of_Purchase	Owner	Fuel_Type	Transmission	RTO	Insurance	Insurance_Type
0	2010 Maruti Alto LXI	₹ 174,699	Abut of 5	3686	Kilometers34,854 km	Year of PurchaseMay 2010	OwnerFirst Owner	FuelPetrol	TransmissionMANUAL	RTOAP09	NaN	Insurance TypeInsurance Expired
1	2013 Maruti Wagon R 1.0 VXI	₹ 333,999	4.2out of 5	3686	Kilometers39,541 km	Year of Purchase,July 2013	OwnerSecond Owner	FuelPetrol	TransmissionMANUAL	RTOAP28	Insurance5/8/2021	Insurance TypeComp
2	2014 Maruti Wagon R 1.0 VXI	₹ 353,199	4.3out of 5	3686	Kilometers23,233 km	Year of PurchaseMarch 2014	OwnerSecond Owner	FuelPetrol	TransmissionMANUAL	RTOAP28	Insurance28/4/2021	Insurance TypeComp
3	2013 Hyundai Eon D LITE PLUS	₹ 257,899	4.4out of 5	3686	Kilometers27,749 km	Year of PurchaseNovember 2013	OwnerFirst Owner	FuelPetrol	TransmissionMANUAL	RTOAP13	Insurance1/1/2022	Insurance TypeComp
4	2017 Hyundai Eon ERA PLUS	₹ 300,699	4.4out of 5	3686	Kilometers12,236 km	Year of PurchaseAugust 2017	OwnerSecond Owner	FuelPetrol	TransmissionMANUAL	RTOTS10	NaN	Insurance TypeExpired

Afterrawdatahasbeencollectedandstoredtol ocal database, The primary Raw data contains 32158 sampleswith12featuresattributeslikeName,Price, Rating, City, Kilometers, year of purchase, owner, fuel type, rating of the car, RTO, insurance and insurance type. Raw data cleaning step was applied usingdifferentfunctionanddifferenttechniqueslike spilt function, loops concepts and rawdata converted into useful data which attributes contains Name of the model, Company of the car, year and month of purchased, kilometers travelled, owner type, transmission, fuel type, city, RTO insurance dataandtype.Afterthesecondarystagedataislooks like shown in afigure

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 8 Aug 2021, pp: 1135-1145 www.ijaem.net ISSN: 2395-5252

	Price	Rating	city	Kilometers	Owner	Fuel_Type	Transmission	RTO	Insurance	Insurance_Type	year	month	Company	Model
0	174699	4	3686	34854	First	Petrol	MANUAL	RTOAP	NaN	Expired	2010	5	maruti	alto lxi
1	333999	4.2	3686	39541	Second	Petrol	MANUAL	RTOAP	6/8/2021	Comp	2013	7	maruti	wagon r 1.0 vxi

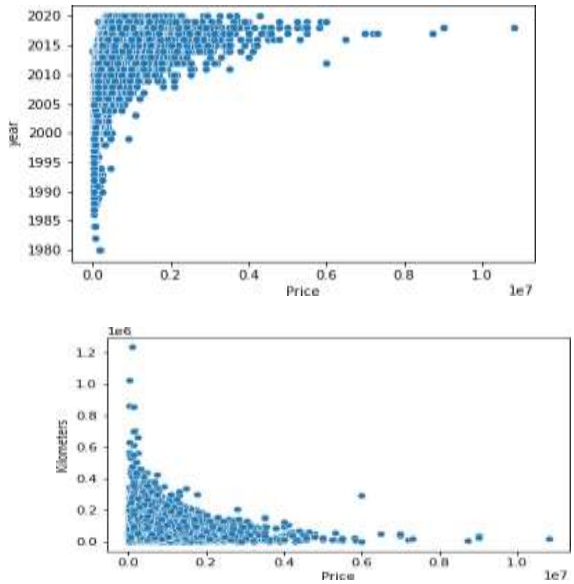
V. EXPLORATORY DATA ANALYSIS:

Afterrawdatacleaning,thenthenextforemost step is Exploratory Data Analysis (EDA) to get the data understanding of each column and how the column is useful for predicted the model. On a careful analysis, many of the features were sparse and they do not contain useful information for prediction.

Hence,itisdecidedtoremovethemfromthedataset.

The attributes RTO and city were completely removed. After the EDA data set contains 32158 samples with 12 features Next, with the data set were perform different statistical operations like Mean, Median, Mode, Count, frequency, maximum, minimum, Quartile 1, Quartile 2, Quartile 3, Quartile 4, Skewness and Kurtosis and standard deviation.

Visualization: To check whether the distribution follows any linearity or not



The above graph shows that the main attributes are not following any linearity between them.

VI. DATA CLEANING:

AfterEDAourdataisinperfectlymanipulated now, in this data cleaning process we were mostly deals with Missing value and outliners treatment. On a careful data analysis, we find that some of the attributes are like Transmission, Insurance, kilometers and type of insurance have missing values. Transmission is a categorial variable so, we were used different techniques like Hot Deck Imputation, Simple imputer and IterativeImputer.

Hot Deck Imputation: It is a method for handling missing values in which each missing data is replacedwithanobservedresponsefroma"similar" unit. It involves replacing missing values of attributes with observed values from a respondent (thedonor)thatissimilartothenon-respondentwith respect to characteristics observed by bothcases.

Simple Imputer: Simple Imputer is a scikit-learn class. It is one of the most useful technics to handling the missing data in the predictive model dataset. It replaces the Nan values with a specified placeholder. It is implemented by the use of the Simple Imputer () method which takes the arguments like missing values, fill value and strategy.

Iterative Imputer: In this process where each feature is modelled as a function of the other features, where missing values are predicted using different regressor like Kenn, Extra tree, Bayes ridge and Decision tree. Each feature is imputed

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 8 Aug 2021, pp: 1135-1145 www.ijaem.net ISSN: 2395-5252

sequentially, one by one, allowing prior imputed values to be used as part of a model in predicting subsequent features.This iterative process is repeated multiple times, allowing ever improved estimates of missing values to be calculated as missing values across all features are estimated. This is also known as fully conditional specification (FCS) or multivariate imputation by chained equations (MICE).

For categorial feature like Transmission, insurance and type of insurance we were used Simple imputer with mode and continuous variable like kilometers we were used simple imputer with mean, Hot deck imputation and iterative imputer compared the results with skewness. Overall, by the analysis Extra Regressor gives us a wonderful skewness with—

After solving the puzzle of Missing values, we here with biggest influencing factor is outliners. Clearing outliners is the heart for feature Engineering because this can influence the distribution and effect the model. There are different techniques to detect and handle outliners. To detect the outliners we were used Boxplot, Inter quartile range, Grubb's test, Isolation Forest and scatter plots. Treating outliners is the most important aspect in feature engineering we were used Fissurization, Arbitrary Outlier Capper and rectify using Inter quartile range.

Treatment of Outliners:

Fissurization: It is a method of averaging that initially replaces the smallest and largest values with the observations closest to them. It changes distribution to gaussian distribution.

Arbitrary Outlier Capper: Using Arbitrary capping methods we can cap the distribution with maximum and minimum values to get better output.

Setting Quartile range: Dealing missing values with quartile range gives a major impact of the outliners. Outliners will remove completely and replaced with maximum and minimum values. They were different kinds of ratio were given to remove the outliners they were 90-10, 0.75-0.25.

Dropping of Outliners: we check the percentage of outliners and drop the outliners to get maximum benefit.

We used above all methods and get the maximum output with setting quartile ranges. After the Data cleaning Process now, we don't have any outliners and missing values in the data now, we can go for

Detection Methods: Using Z- score or Extreme value analysis to detect the unanimous behavior in the dataset and this indicates how many standard deviations a data point is from the sample's mean, assuming a gaussian distribution. Using Z score I detected that in Kilometers travelled and rating attributes have an outliner.

Box-plot: Using Boxplot we were used to detect the skewed of the distribution and to check were the outliners present in the distribution.

Inter- quartile range: By calculating the inter quartile range of attributes in the data set I find that some of the data points are highly influencing for the dataset.

Grubb's test: Using Grubb's test and p value to find the alternative hypothesis and null hypothesis.

Isolation forest: The principle beyond the Isolation Forest algorithm outliners is few and far from the observations.

Using all the above methods, we were checked and come to conclusion that they are two outliners' attributes present in the dataset which is kilometers travelled and rating given.

data wrangling.

Data Wrangling: Basically, Data wrangling is divided into Encoding, Binning, train-test spilt, scaling and transformation.

Transformation: we were checked every attribute of the dataset and distribution of the transformation and we comes to know categorical variable are following Poison and Binomial distribution and continuous variable having the skewness greater than 1 and less than -1. So, we given log normal distribution to get skewness greater than 1 and less than -1.

Train-Test spilt: Using train test spilt we divide the data into 2 parts with 80% of the train and 20% of the test data. Our total data consists of 32158 records. In which 80% means 25726 records for training and 6430 records for testing.

Encoding: For both the train and test data we perform different types of methods to convert the data from categorical to numeric so, machine learning algorithms can easy to proceed. We used

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 8 Aug 2021, pp: 1135-1145 www.ijaem.net ISSN: 2395-5252

One hot encoder, dummy variable trap, label encoder for the data. This all are one type of method and another method we just keep the categorical feature as same because it is useful for cat boosting algorithm.

One hot Encoding: This method is primary use for nominal data in which there is no comparing between the columns. It is a representation of categorical variables as binary vectors. This first requires that the categorical values be mapped to integer values. Then, each integer value is represented as a binary vector that is all zero values except the index of the integer, which is marked with a 1.

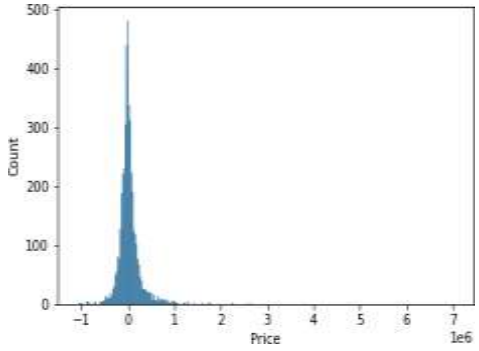
Dummy Variable: It is also known as indicator variable. This method feeds the data to machine to understandthebehaviorofcategoricalvariableand gives the input in the values 1 or2.

Label Encoder: It is useful for to encode the ordinal data to feed for machine. Using Label encoder, I divide the attributes like company and model. The attributes like Transmission, Owner and fuel type convert using one hot encoding and dummy variable.

Feature Transformation: Attributes like Ratings and Kilometers are not following standard Normal distribution So; we can convert This attribute using log Normal transformation and Box cox transformation.

Modelling: We used different Algorithms like Multiple linear Regression, Random Forest, Gradient Boosting Machine (GBM) Regressor,light Gradient Boosting Regressor, Extreme Gradient Boosting Regressor and cat boost Regressor and Adaptive Boosting Regressor

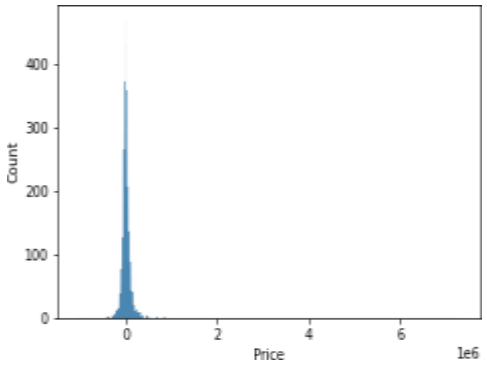
1. Multiple Linear Regression: Multiple Linear Regression was chosen as the first model due to its simplicity and comparatively small training time. We haven't used any regularization since results clearly showed lowvariance.



International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 8 Aug 2021, pp: 1135-1145 www.ijaem.net ISSN: 2395-5252

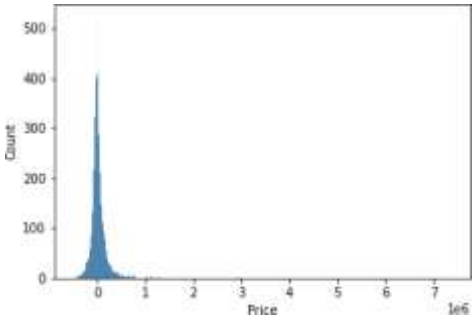
The above histogram shows the errors of the predicted variable. This clearly states that it isright skewed distribution. The more errors are closed to Zero and minimum of error inthe range between -1 to1.

Random Forest Regressor: Random Forest Regressor is an ensemble learning based on decision tree model. Specifically, this Random Forest is designed by Ho to reduce the overfitting problem of Decision Tree. A random forest algorithm consists of many decision trees. The 'forest' generated by the random forest algorithm is trained through bagging or bootstrap aggregating. Bagging is an ensemblemeta-algorithm that improves the accuracy of machine learning algorithmsThismodelwashencechosentoaccount for the large number of features in the dataset and compare a bagging technique with the following gradient boostingmethods.



Theabovehistogramexhibitsthepredictivev ariable errors. This clearly states that the distribution is Right Skewed. Most errors are closed to Zero with minimal error are less than0.25.

Gradient Boosting Regressor Machine: Gradient Boosting is another type of decision tree-based method that is generally described as "a method of transforming weak learners into strong learners". This means that like a typical boosting method, observations are assigned different weights and based on certain metrics, the weights of difficult to predict observations are increased and then fed into another tree to be trained.; it allows for the optimization of arbitrary differentiable loss functions.Ineachstagearegressiontreeisfitonthe negative gradient of the given loss function. This, allows for the optimization of arbitrary differentiable loss functions. In each stage a regression tree is fit on the negative gradient of the givenlossfunction.Thismodelwaschosento account for non-linear relationships between the features and predicted price.

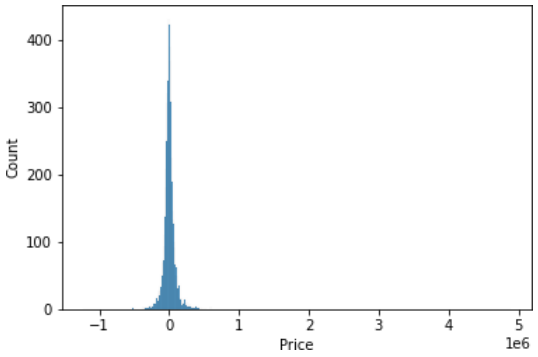


Theabovehistogramexposesthepredictivevariable errors. Most errors are closed toZero.

Extreme Gradient Boosting (XGBoost): XGBoostis quite similar to the gradient boosting algorithm but features many additive features that significantly improve its performance such as built-in support for regularization, parallel processing as well as giving additional hyperparameters to tune

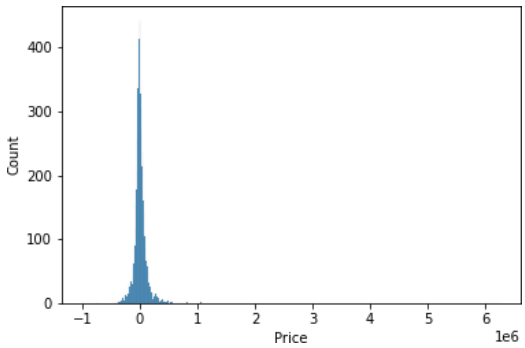
International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 8 Aug 2021, pp: 1135-1145 www.ijaem.net ISSN: 2395-5252

such as maxdepth, subsample and sampling method. A maximum depth of 16 was used and the algorithm was run on all cores inparallel.



The above histogram shows predictive variable errors.ThisclearlystatesthatPeakness(Kurtosis)is higher and it is right skewed Skewed. Most errors are closed to Zero. This is best Comparedto

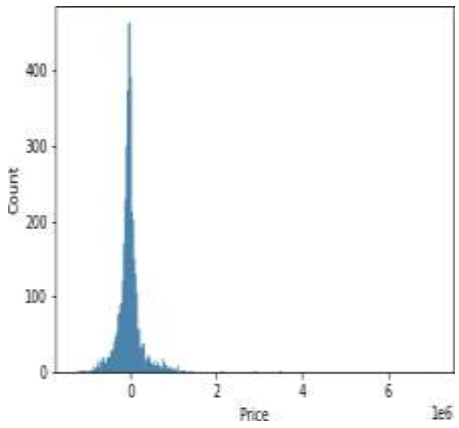
Light GBM : Light GBM is another gradient boosting based framework which is gaining popularity due it higher speed and accuracy compared to XGBoost or the original gradient boosting method. This Light GBM has a leaf-wise tree growth instead of a level-wise approach resulting in higher loss reduction.



The above histogram displays the predict variable error. It is clearly Visible that the range of errors is less than 0.5 to 0.5

Adapative Boosting Machine: It is called Adaptive Boosting as the weights are re-assigned to each instance, with higher weights to incorrectly classified instances. Boosting is used to reduce bias as well as the variance for supervised learning. It works on the principle where learners are grown sequentially. Except for the first, each subsequent learner is grown from previously grown learners. In simple words, weak learners are converted into strongones.

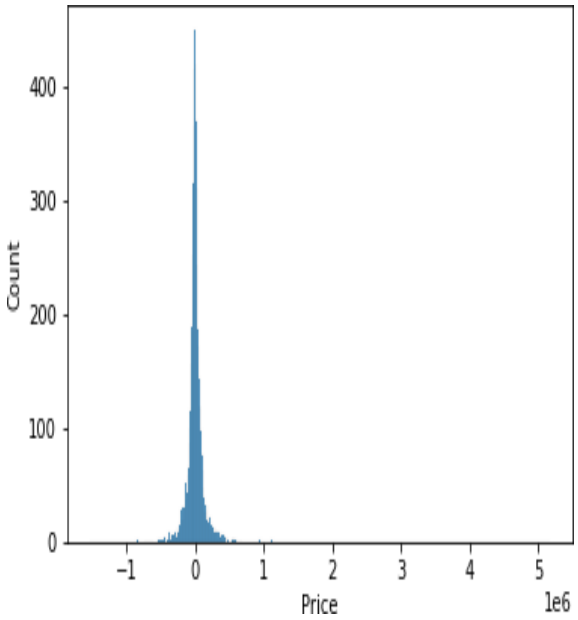
International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 8 Aug 2021, pp: 1135-1145 www.ijaem.net ISSN: 2395-5252



The above histogram demonstrates the predict variable error. It is clearly Visible that the range of errors are very minimal

The overhead histogram describes the predict variable error. It is clearly Visible that the range of errors is minimal to zero.

Cat Boost Regressor: Cat Boost supports various types of data categories like numerical, categorical and text**.** The special feature with this categorical data will handle directly without using encoding techniques and having many features to handle. In Cat Boost "Boosting" is refers to the **gradientboosting machine learning**



The above histogram displays the predict variable and it error of the predicted variable. It is Visible that the peakness is achieved at 400 and the range of error is between -1 to 1. The special about this will deal with categorical variables as it is and not required any kind of change like encoding.

Coefficient of Determination (R^2): The coefficientdetermination is one of the evolution metrics for Regression that examines how differences in one variable can be explained by the difference in a second variable, when predicting the outcome of a given event. This can be denoted as R^2.

Evaluation Metrics:

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 8 Aug 2021, pp: 1135-1145 www.ijaem.net ISSN: 2395-5252

R^2= 1- SSres / SStot

Where,

SSres is the sum of squares of the residual errors.

SStot is the total sum of the errors.

Root Mean Squared Error: The most commonly used metric for regression tasks is RMSE (root-mean-square error), also known as RMSD (root-mean-square deviation). This is defined as the square root of the average squared distance between the actual score and the predicted score.

Here, yi denotes the true score for the ith data point, and yi Denotes the predicted value. One intuitive way to understand this formula is that it is the

$$\text{RMSE} = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{n}}$$

Euclidean distance between the vector of the true scores and the vector of the predicted scores, averaged by n, where n is the number of data points.

VII. RESULTS:

The results of our test are as follows

Name	Train(R2)	Test(R2)	Train (RMSE)	Test (RMSE)	Remarks
Linear Regression	0.719	0.727	0.449	0.444	Low R2
Random Forest	0.987	0.916	0.096	0.243	Over Fitted Model
Gradient Boosting	0.867	0.879	0.308	0.295	Moderate Model
Extreme Gradient Boosting	0.980	0.944	0.118	0.120	Good model
Light Gradient Boosting	0.923	0.925	0.234	0.232	Good Model
Cat Boost	0.948	0.922	0.192	0.192	Good Model
Ada Boost	0.660	0.655	0.502	0.506	Low R2

VIII. CONCLUSION:

The results shows that Compared to Linear Regression, most Boosting methods performed comparably well. We get high variance with the ExtremeGradientBoostingMachineandveryslight less with Light gradient boost, Gradient boosting and cat boost machine. The extreme gradient gives train r2 score with 0.980 and with test 0.944 which good and optimized model for predicting thesecond carprice.

FUTURE WORK: For better performance, we plan to judiciously design deep learning network structures with Artificial Neural Networks and train on clusters of data rather than the whole dataset and to use different procedures to select the foremost features for dataset

REFERENCES:

[1]. Listiani, M. (2009). Support vector regression analysis for price prediction ina car leasing application (Doctoral dissertation, Master thesis, TU Hamburg-Harburg)
[2]. https://www.temjournal.com/content/81/TEMJournalFebruary2019_113_118-Journalon CarPrediction using Classification.
[3]. sklearn.ensemble.RandomForestRegressor— Retrieved from:https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html
[4]. Catboost regressor-https://catboost.ai/docs/concepts/python-reference_catboostregressor.html
[5]. car24.com:https://www.cars24.com-Data retrieved fromExtraGradientBoosting-https://xgboost.readthedocs.io/en/latest/
[6]. Aprojectreportoncar24.comhttp://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/26612934
[7]. Pudaruth, S. (2014). Predicting the priceof used cars using machine learning techniques. Int. J. Inf. Comput. Technol, 4(7),753-764.