PROJECT REPORT

Course title: Big Data and IOT Lab

Course Code: CSE 413

Project Title: Used Car Price Prediction

Submitted to:

Mr. Amir Sohel
Senior Lecturer
Computer Science and Engineering

Submitted by:

Name: Md. Ashfak Ullah Nafis (Report writing)
ID: 201-15-14110 Sec: G_55
Name: S.M. Mahamudul Haque (Coding)
ID: 201-15-13707 Sec: G_55

Used Car Price Prediction

1.Introduction:

Car is not just a machine which we use in our daily life to comfort us. It is an emotion. People want to buy a car, not only for safe and comforted travel, sometimes they have a dream and vision to buy their own car. But some people can't afford to buy a brand new car and that's the reason they choose to buy the used cars. Our research is basically offers information on used vehicles are value for money or not, what they are sold for, and all the specifics about their condition. It is a thorough collection of useful data regarding used automobiles. There has never been as much demand for precise and trustworthy techniques to forecast used car prices given how quickly the automotive industry is changing. Advanced predictive models have been developed as a result of the dynamic factors influencing the value of pre-owned vehicles and the growing reliance on data-driven decision-making. The cost of a used car is a complex puzzle with many moving parts. These include the make, model, and mileage of the car as well as outside variables like market trends, local preferences, and economic situations. The combination of data science and automotive knowledge has made it possible to develop novel approaches for used car price prediction in recent years. Through the utilization of past sales information, industry patterns, and innovations in technology. While there are obstacles in the way of accurate used car price prediction, like market volatility and the requirement for large and diverse datasets, there are also creative solutions that can be found. When combined with in depth knowledge of the automotive industry, advanced modeling techniques can enable stakeholders to make well informed decisions when purchasing, disposing of, or determining the value of used cars. Accurately predicting used car prices becomes a transformative tool as we navigate the intersection of automotive commerce and technology. We set out on a mission to unravel the secrets of the used car market by utilizing data science, machine learning, and real-time market insights. Along the way, we will offer priceless advice to both industry participants and enthusiasts. In this quest, what lies ahead holds not only forecasts but also a more precise understanding of the complex dynamics influencing the worth of our reliable four wheeled allies. And this important parts are forced us to work on this type of datasets.

2.Methodology

Our primary goal of this research is to analyze, investigate, and develop knowledge, insights, or understanding about this topic, issue & occurrence. Depending on this method and serves numerous critical goals. Our goal in this research is to add to the body of knowledge by learning new information.

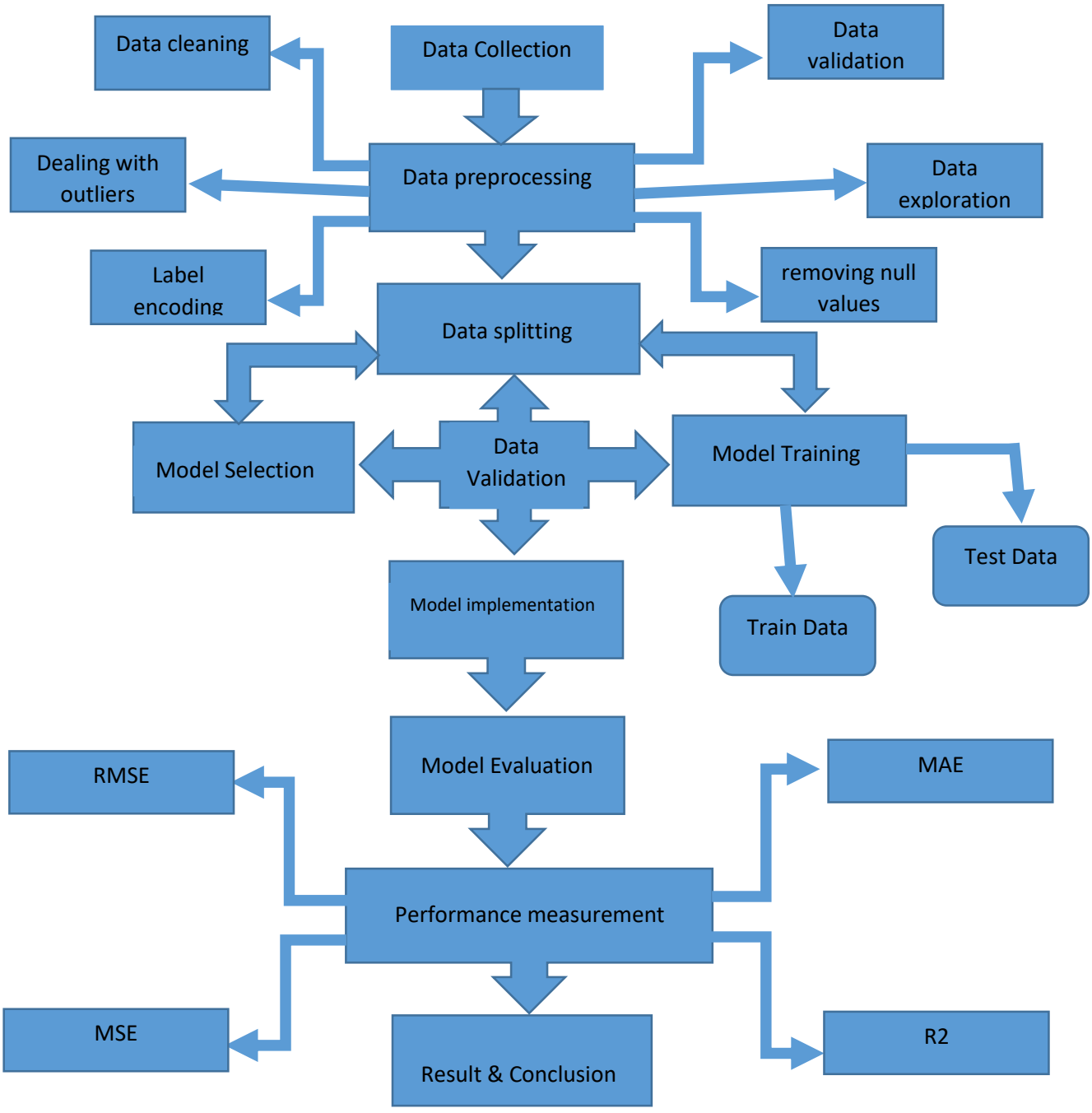

figure: proposed model for used car price prediction

We also plan experiments or studies to see if particular hypotheses or predictions are supported by implementing this data into machine learning, which produces data that can be used to support or disprove claims, arguments, or propositions.

Logistic regression: Basically people work on two types of data one is image processing and another is statistical data. We work on statistical data, and we know Logistic Regression is a statistical method and a key algorithm in machine learning that is used for binary classification tasks. So, we try to make it by the process which can also be used as the logistic [2] function in our research, as this regression method is used to model the connection between a binary dependent variable and one or more independent variables so we use it and frequently apply it to our dataset.

Random Forest Regression: As our dataset carries a lot of data so we need to represent our data on some great graphical method. So, we use this machine learning approach that expands the application of random forests from classification to regression issues. Because it helps our data as a flexible and effective ensemble learning technique that integrates various decision trees to produce precise predictions on continuous (numeric) target variables. Numerous fields, including finance, modeling, frequently use Random Forest Regression that shows our data like a tree model.

Imbalanced Data: It is hard to balance our whole dataset especially when the classes are imbalanced, our data set combined with one class exceeds the other, and we know that the categorization by binary can be difficult. So, we apply the techniques and the use of appropriate evaluation metrics that can aid our data as treatment of imbalanced datasets.

By helping with these methodical and structured techniques to perform our research, problem-solving, fixing, or carrying out numerous operations in an orderly and rigorous method. It simplifies our project planning and execution by breaking down complex procedures into simple steps. And also help us to discover problems, analyze root causes, and devise solutions. By ensuring the research method is conducted in an orderly way that can be examined by peers for validity and reliability. This is especially significant in sectors such as car management and manufacturing, where improving methods is essential.

3. Dataset Description:

This dataset has a starting point for investigating the variables that influence used automobile prices. With the help of elements like name, seller, price, among other data points, it is possible to learn how various aspects of a vehicle affect how much used automobiles cost. Car is not just a machine which we use in our daily life to comfort us. It is an emotion. People want to buy a car, not only for safe and comfortable travel, sometimes they have a dream and vision to buy their own car. But some people don't afford to buy a brand new car and that's the reason they choose to buy the used cars. It offers information on how used vehicles are sold, what they are sold for, and all the specifics about their condition. It is a thorough collection of useful data regarding used automobiles. Each advertisement includes details like date car's name, the seller's status (private or dealer), price, vehicle type, year of registration (Type of vehicle and at which year was the car first registered), gearbox type, power output in PS (horsepower), model of car, how many kilometers has it driven so far & fuel type utilized by it (petrol/diesel /electricity/ lpg).

About dataset:

Our dataset is about car price prediction. If we want to do any experiment on data, we must to collect a great meaningful and valid dataset. So, firstly we search for real life dataset. But it is hard to get those type of data. Because people got refuse to give their data because they felt that they will face security issues if they share those data. So, we search online for dataset and finally find a dataset on Kaggle. And that data is decent for our work. Here is the dataset link:

https://www.kaggle.com/datasets/thedevastator/uncovering-factors-that-affectused-car-prices

And this dataset contains Around 372000 rows and 20 columns. So, we can consider this dataset as a big data. And this data is suitable for our work.

Dataset description table

S. No.	Column name	Description	Data Type
01	**Price**	It represents how much amount of that car or how much money we need to spend to buy this car. And it represents as a number.	Integer
02	**vehicleType**	Type of vehicle explains the categories of vehicle that people use in their regular life according to their term of use.	String
03	**yearOfRegistration**	Register year means which time the car has been registered. The year that a vehicle was produced by the carmaker is referred to as the automobile registration year, often referred to as the model year or production year.	Integer
04	**gearbox**	A car's gearbox refers to the transmission system that controls the transfer of power from the engine to the wheels. There are various types of gearboxes used in automobiles, each with its own set of qualities and advantages. There are two type of gearbox (manual or automatic).	String
05	**powerPS**	Power of the car in PS. PS or PferdStarke is the metric measure of horsepower. It is the equivalent of 98.6% of one HP.	Integer
06	**Model**	There are so many brand cars in the world. And there are also so many categories on it, a manufacturer's particular version or variant of an automobile is referred to as a "model" in the automotive industry.	String

07	**kilometer**	It refers to the distance the vehicle has traveled in total since it left the dealership from the present time, Automobile's odometer can be used to calculate its mileage in kilometers	Integer
08	**fuelType**	Fuel type describes the particular kind of fuel that a car or other machine uses to power its engine. The type of fuel used can have a big impact on how well	String
		a vehicle performs, how much it costs to operate, and so on.	
09	**Brand**	A car brand symbolizes the manufacturer's identity and reputation and is used to identify their automobiles from those of other firms. Each automobile brand has its own emblem, design philosophy, and model lineup that caters to different market niches and consumer preferences.	String
10	**notRepairedDamage**	Because there are so many vehicles on the road, multiple crashes occur and cars get damaged. However, a car needs to be repaired or recovered after being used for so long. Whether or not the car has any damage that has not been repaired.	String

4.Data Preprocessing:

The pipeline for data analysis and machine learning must include the collection of data as an essential phase. In order to gather raw data for analysis or for training machine learning models, it must first be cleaned, transformed, and organized. The purpose of data preprocessing is to increase the data's usefulness and quality, making it easier to deal with and guaranteeing that any analysis or modeling outputs are more accurate and understandable.

Data cleaning:

When we download the dataset we saw that there are so many row and columns which are around 372000 rows and 20 columns. But all of these rows and columns are not suitable for work, that is the reason we need to clean so many rows and half of the total columns. After cleaning the dataset, we find that 101019 rows and around 10 columns and these are great for our work or we can say these data's are actual data.

Removing null values:

we write a code to show that what amount of null values are held in a column or a column carries how much null values. So, applying that code we see that vehicle Type (10185), gearbox (5452), model (5545), fuel Type (9137), not Repaired Damage (19539) these columns carries null values with that amount of numbers which is on brackets. In this figure the white lines are the null values. Then we apply percentage method all null values. And we see that the highest amount of percent is 19%. Here we apply a condition if null values are more than 25% so the column should be dropped, but our highest percent is 19% and that is the reason the column are not dropped. After that we start to dropping rows, after dropping rows our and removing null values we find 71133 rows with actual values.

Dealing with outliers:

Here at first we describe the data and it shows so many attributes like min, max, 25%,75%. In the price section we put the value 0.9 and it means it actually contains 90% of the graph. And it basically counts from left side to right side. And if we take 0.9 on the price quantile we can see the highest amount of price is 15980. After updating the data again, we find the min and max price. It shows min price is 0 and max is 15980.As we know that no one can buy or sell a car on 0 dollar and for that reason we take 0.3 and now we can see on the graph that the min price is 250 dollars and max 15980. So, we can now work on this data. Same technique has been applying on the year of registration, here we cut down 10% from the left side. The minimum year is 1997 and maximum year is 2018. Now let's come to the power ps, here we apply 0.34 and the min power ps is 22 and max is 400. After applying the same technique car kilometer, we finally find that the amount of valid data in our data set is about 52349 from101019.

Feature engineering: The exact term is when we can see the great graphical representation and in our dataset it is essential to see that term in order to identify

the most illuminating characteristics in our paper it is our particular situation, feature engineering is an inventive and iterative process in which we test out various transformations and combinations. It frequently plays a similar role to choosing the best algorithm and fine-tuning hyper parameters in determining how well our machine learning models perform and do their terms perfectly.

Quantile Method: We use this quantile method because we know that the quantile method, also known as quantile regression, is a statistical method for estimating conditional quantiles of a dependent variable, revealing details about how the relationship with independent variables varies at various points in the distribution. When handling non-normally distributed data or data that contains outliers, it is especially helpful to create our data to look more simple.

Label encoding:

There is a dictionary library we need to input in our code and it trying to labeling the text column in our data set there is a column which is gearbox type so here are two types of gearbox automatic and manual so it gave them the separate number. We need to input this dictionary library on every text column and it comes output as a separate number. We can easily see this into the figure. We also applied this method to these columns: vehicleType, Model, fuelType, Brand, notRepairedDamage to transform the text data into numerical data.

Data validation: According to our dataset we can confidently say it is a perfect data, because we can work in this data set without any kind of problem. Because to perform the action we need to apply a valid dataset. So, at first we examine data to ensure its accuracy, consistency, and compliance with established rules or standards. Our main goal is to make sure that the data is accurate, trustworthy, and appropriate for the use to which it is being put. We also make sure when we do the terms like data entry, data import, data preprocessing frequently, involve data validation, which plays an important role while we use all of these components of data quality management.

Forecasting and prediction: Prediction is the thing that we want to ensure in our whole research method that is the reason we work on applying lag features. That plays an important part in time series analysis and forecasting, analysts and data scientists. It may use previous other research information that creates better

predictions. And here we analyze data trends, and account for temporal relationships to maintain The particular lag values and number of lag features utilized are determined by the nature of the data and the goals of the analysis of our research.

List wise Deletion (Row Removal): After some preprocessing method we need to handle missing data in a dataset by deleting all of the rows with missing values. For that reason, the entire row is excluded from the analysis and it is easy to see if any of our data points have certain row and missing values, and then our data analysis results may be significantly impacted by this simple, and our data strategy is also easy to implement.

Imputation for Regression: Using a regression model here is to anticipate the missing values, regression imputation is a method for dealing with missing data. Because this data set carries lots of data which considers the relationships between variables in the dataset, we need to approach straightforward imputation techniques like mean or median imputation in our research.

5.Feature extraction:

Feature is actually representing the independent values and here are so many independent values in our dataset which are vehicle Type, year of Registration, gearbox, power PS, Model, Kilometer, fuel Type, Brand, not Repaired Damage. These are not depending on any other attributes. In our dataset there is only one dependent value which is Price, because we need to figure out the price and working on it. So, we declare price as a Y value and the rest is X value. Then we split it on train and test data.

5.1 Label encoding:

There is a dictionary library we need to input in our code and it trying to labeling the text column in our data set there is a column which is gearbox type so here are two types of gearbox automatic and manual so it gave them the separate number. We need to input this dictionary library on every text column and it comes output as a separate number. We can easily see this into the figure. And for better result we introduce this valuable term to express our work. Although it is the most important term.

5.2 PCA:

A popular dimensionality reduction method in data analysis and machine learning is principal component analysis, or PCA. Its purpose is to maintain as much of the variability of the original data as possible while simplifying and representing complicated, high-dimensional data in a lower-dimensional space. In our dataset here are already have some selected number of components. Which represents that how much columns we actually want, as an example we can say we gather 9 columns each other and make 1 column. And after implement that we find the result 1. But after implement the PCA our result is getting even worse.

5.3 Feature Scaling:

To normalize or define the range of independent variables or features from our dataset, we have been used feature scaling is a preprocessing technique frequently in machine learning and data analysis. In order to ensure that machine learning algorithms function well, feature scaling is crucial because many of them are sensitive to the input feature scale. The requirements of the machine learning algorithm we are using and the unique properties of our data will determine which scaling method is best. Standardization is typically a good option when you want to maintain the original distribution's shape and make the features more amenable to specific algorithms, such as gradient-based methods. On the other hand, min-max scaling is helpful when we want to constrain the feature values to a specific range. And we include this term to get the better result for our experiment.

6.Splitting train and test data:

If we want identify the result with model or without model and see this two type of compression, we must apply the method splitting data. First, splitting the data into 80% training and 20% combined test and validation, after that splitting the remaining 20% into 50% test and 50% validation and at last Verifying the shapes of the datasets. And fulfill the condition which is 80% train data 10% test data and the rest of 10% is validation data. After applying the model, we actually need the array. And the output is shown as:

X_train shape: (41879, 9), y_train shape: (41879,)

X_test shape: (5235, 9), y_test shape: (5235,)

X_val shape: (5235, 9), y_val shape: (5235,)

7.Performance Measurement Matrices:

In many disciplines, especially those involving machine learning, data analysis, software development, and quality control, accuracy testing is a crucial part of assessing how well models, algorithms, or systems function. In our data set we use 11 accuracy testing technique. Which now I am going to explain in below.

i. R_2 score:

In this graph We can see GBR is contain around 0.84 of this graph second is neural network regulation which contain 0.83 of the total graph and also we can see the KNN which contains 0.872 of the graph and the next is decision tree regression it contains 0.966 of the total graph but the highest number of the graph is random forest which contains 0.95 of the graph. so we can say that random forest is the highest of this graph.

ii. Mean square error:

Now let's go to the second part which is mean square error. At first we can see in the graph, GBR contains 240 on the graph. and the second number of the graph is neural network regression but the highest amount of the draft is linear regression which contains 4390 265 of the graph and this is the highest of this sector. and the lowest part of this graph is decision tree regression.

iii. Root Mean square error:

In this graph the highest number contains 2095 which is linear regression, and the second highest of the graph is NNR which contains 1729 of the graph. the lowest number of the graph is the decision tree. Second lowest of the graph is random forest, and that is the summary of the graph.

iv. Mean absolute error:

Here in this graph the perfect result is shown by the linear regression and after that NNR is also showing the great result in this graph. These two are Create a role of this entire graph and others also share their content amount of this graph.

v. Mean squared log error:

KNN is the highest number of peace graphs, and it contains 0111 of this graph and the second of these graphs is random forest which contains 0.051 in this graph. and the other one is null value which contains zero in this graph.

vi. Mean absolute percentage error:

The highest number of this graph is linear regression and the second highest of the graph is KNN number third is going to be random forest. so it is surely provided that the first one contains around 0.68 of the total graph and their difference between all of this group.

vii. Median absolute error:

The lowest number of the graph is decision tree regression when the second lowest in this graph is random forest. The highest number of the graph is linear regression which contains 12711.73 all the graph. and The second highest of the graph is KNN, which contains 581 of the graph.

viii. Max error:

Max Error is a useful metric, but to get a more complete picture of a model's performance, it's usually combined with other metrics like mean absolute error, mean squared error, or R-squared.

ix. Explained variance score:

An easy way to evaluate how well a regression model captures the variation in the dependent variable is to look at the Explained Variance Score. But in order to get a complete picture of the model's performance, it needs to be used in concert with other metrics and factors.

x. Pinball Loss:

The highest number of the pin ball loss model is linear regression and the lowest number of the pinball loss is Decision tree regression.

xi. D^2 absolute error score:

In this graph we can see that models are shown on the graph. The highest number of the graph is decision tree regression and the second highest number of the graph

is random forest and Kane is the third highest number of this graph and this easily explains that the term that we use is valuable for it.

8. Data Validation:

As we have got a condition about data validation which is 10% data is consider as data validation among our dataset. According to our dataset we can confidently say it is a perfect data, because we can work in this data set without any kind of problem. Because to perform the action we need to apply a valid dataset. So, at first we examine data to ensure its accuracy, consistency, and compliance with established rules or standards. Our main goal is to make sure that the data is accurate, and we basically work on the condition which is 10% validation on the among dataset and we get some result on that now I am going to show the result in a table.

Name	**validation data result**
r2_score	0.8593049765562055
Mean squared error	2139103.0807382525
Root Mean square error	1462.5672910120247
Mean absolute error	987.4650098984933
Mean squared log error	0.1330376367983777
Mean absolute percentage error	0.30640602339949724
Median absolute error	631.0645238095235
Max error	13517.88538370411
Explained variance score	0.8593093454748271
Pinball loss	493.7325049492467
D^2 absolute error score	0.680665237076548

So, this is the table with the experiment name and validation result.

8.1 Correlation:

Applications for correlation analysis include finding correlations between economic indicators, figuring out how variables affect business metrics, and evaluating the relationships between variables in scientific research. Making decisions based on

data is aided by it, and it can reveal which variables are most important for forecasting or illuminating changes in other variables.

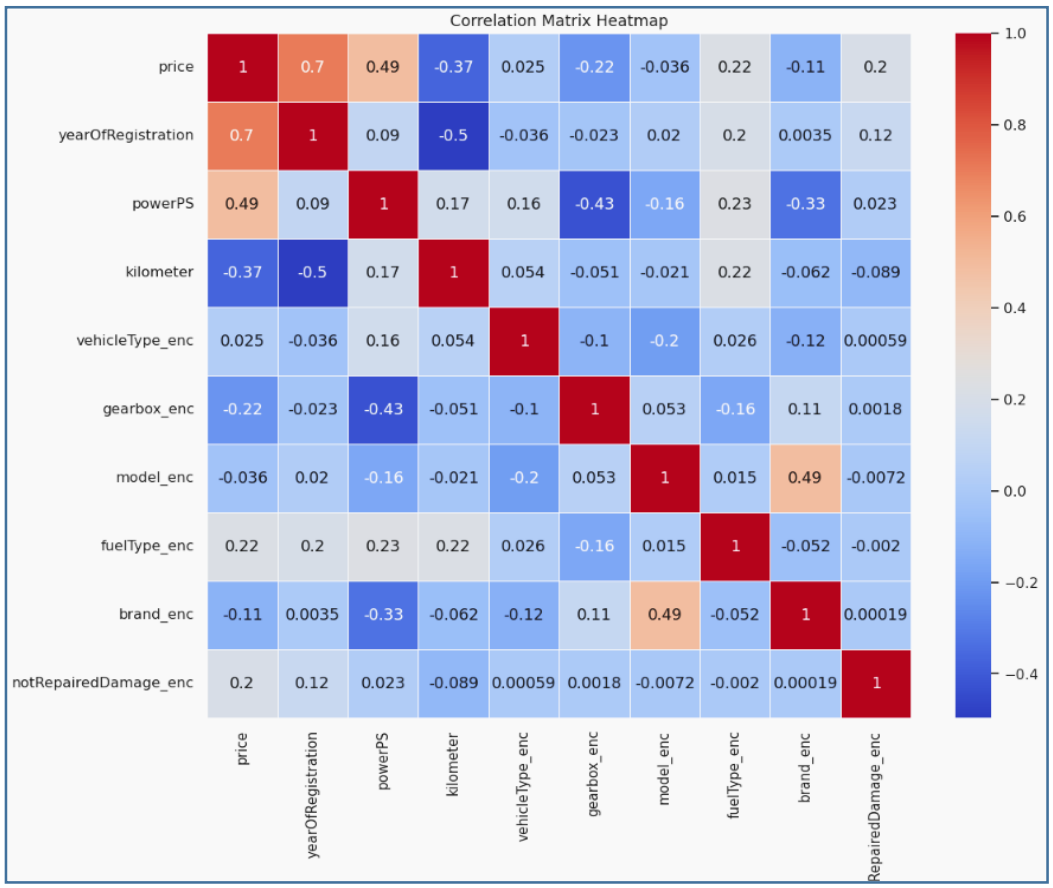

Figure 1: Correlation Matrix Heatmap

We can find patterns and clusters of strongly positively or negatively correlated variables by looking at the correlation heat map. This can aid in our understanding of the underlying relationships in your data and the variables that tend to move in tandem or in opposition to one another. It is imperative to bear in mind that although correlation heatmaps offer significant insights, they are limited to capturing linear connections between variables. This approach might not fully disclose complicated linkages or non-linear interactions.

9. Model implementation

We know that the process of moving a machine learning model from its development and training stages to an operational, useful state where it can make predictions, categorizations, or other data-driven decisions in real-world applications is known as model implementation. And in our experiment we also use this technique for better result and great graphical represent.

Random forest:

We basically use this model to get better result in our research. Widely utilized for our dataset capacity to produce reliable and accurate predictions, we manage noisy and high-dimensional data, and recognize significant features, in that case Random Forest is a flexible and strong machine learning technique. When feature randomization and data sampling are combined with its ensemble technique, it becomes a useful tool for a variety of applications requiring predictive modeling. And we get better accuracy in our output.

Linear Regression:

This is the second model that we have been used in this model implementation. As we know that the basic and popular statistical technique in many domains, including the social sciences, engineering, economics, and finance, is linear regression. It helps our dataset to create predictions, sheds light on the connections between variables, and we present the whole term an easy-to-understand method of data analysis. Because linear regression is strong and helpful, it might not be appropriate for very non-linear relationships, which call for more intricate modeling strategies.

Decision tree regressior: Another model we have use on this term which is decision tree regression While decision tree regression offers certain benefits, like ease of use and readability, it also has drawbacks. If decision trees are not regularized or pruned, they may over fit and become sensitive to slight changes in the data. Tree pruning, minimum samples per leaf, and maximum tree depth are a few common techniques used to control the tree's complexity and enhance its generalization ability in order to address these problems.

KNN: while doing our research we noticed that KNN is a simple algorithm that works well for some datasets, particularly those with a less complicated decision boundary. This is a straightforward and popular technique for machine learning tasks including classification and regression, and we want to use this term into our model

and it's a fantastic place to start if you're new to the field. This bases predictions on the average value of the k-nearest training instances or the majority class.

Neural Network Regression (NNR): An neural network Regression is a kind of machine learning model in which continuous numerical values are predicted using neural networks. We also want to get some better class of result and the Regression challenges differ from classification tasks in that the objective is to translate input information to a continuous output instead of class labels. For problems where the output is a continuous variable, neural networks especially feedforward neural networks work well. They are capable of managing interactions between features and non-linear patterns.

Gradient Boosting Regression (GBR): As we are trying to prove the better result of our experiment so, it is an ensemble learning technique that builds a powerful predictive model by aggregating the predictions from several weak learners, usually decision trees. Gradient boosting's primary principle is to train weak models one after the other, with each one concentrating on the mistakes made by the ones before it. These libraries offer streamlined and effective gradient boosting model training strategies. We use the term for to get some great result and output and it's vital to remember that gradient boosting can be susceptible to overfitting, therefore optimizing performance requires careful adjustment of hyper parameters and regularization strategies.

10.Visual Representation:

The term "visual representation" describes the process of communicating information, data, or ideas through the use of visual elements like charts, graphs, diagrams, and other visual aids. It is an effective way to convey abstract or difficult ideas in a way that makes them simpler for people to comprehend and interpret. The natural processing and interpretation of visual information by the human brain is facilitated by visual representation. It is an adaptable instrument for information transfer, learning enhancement, and creativity cultivation in a variety of contexts and fields. Here we have done some work like train vs test values, plotting the training, checking the prediction with the original values and also plotting testing error.

At first if we want to explain the term which is training time we can see that all the dots or the object that are in visual at the graph is totally gather and close to each other.

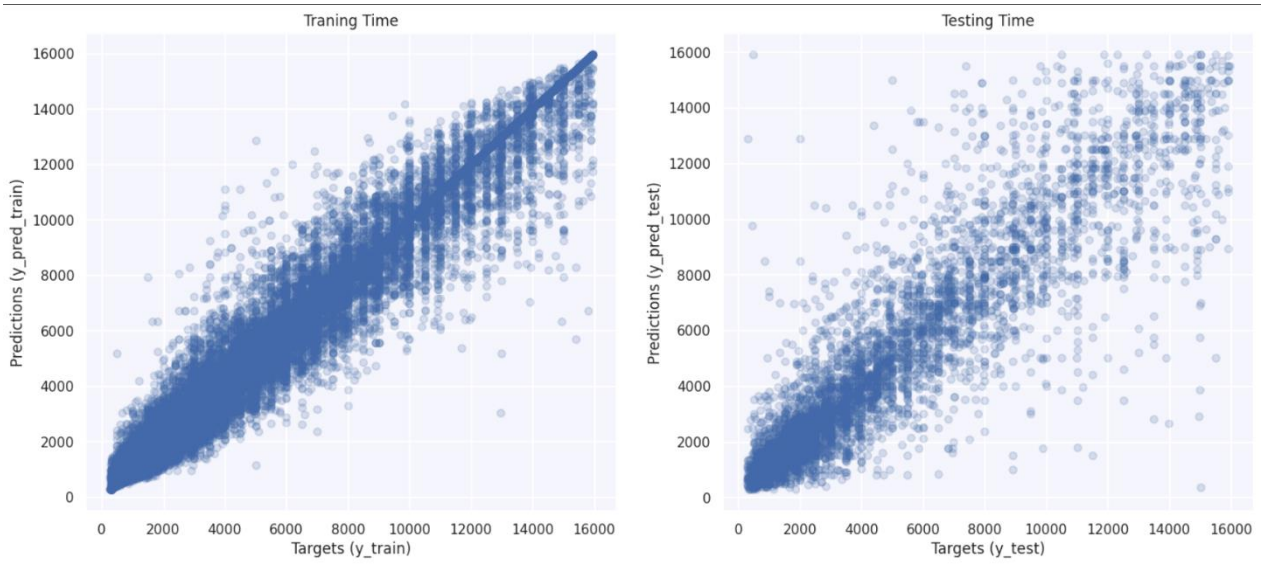

Figure 2: Visual Representation of Training and Testing time

If we notice the Testing time the close and connected dots are distance from each other and the bottom of the graph is 0 and it is the lowest of the graph on the other side the top number contains 16000 of the graph.

10.1Actual vs. predicted values:

A quantity, measurement, or parameter's real or true value as observed, measured, or ascertained in a real-world setting is referred to as its "actual value". A variable or phenomenon's real values are its true or factual state, and they are frequently employed as a benchmark for analysis or comparison. The phrase "predicted value" describes a value that is estimated or forecasted using input data, a set of parameters, or features, and produced by a model, algorithm, or analysis. Given the information at hand, predicted values indicate the course of events or outcome that a predictive model predicts for a given circumstance. Here in the graph the highest number is 16000 and the lowest number is 0. And here also have a middle number of the graph which is 8000. So many plots are in the graph which define the predicted value and the actual value. At the first picture it is not clear that the graph exactly represents

what term, so in the other picture it is define as an ideal value and it is marked with red line and the others are basically plots.

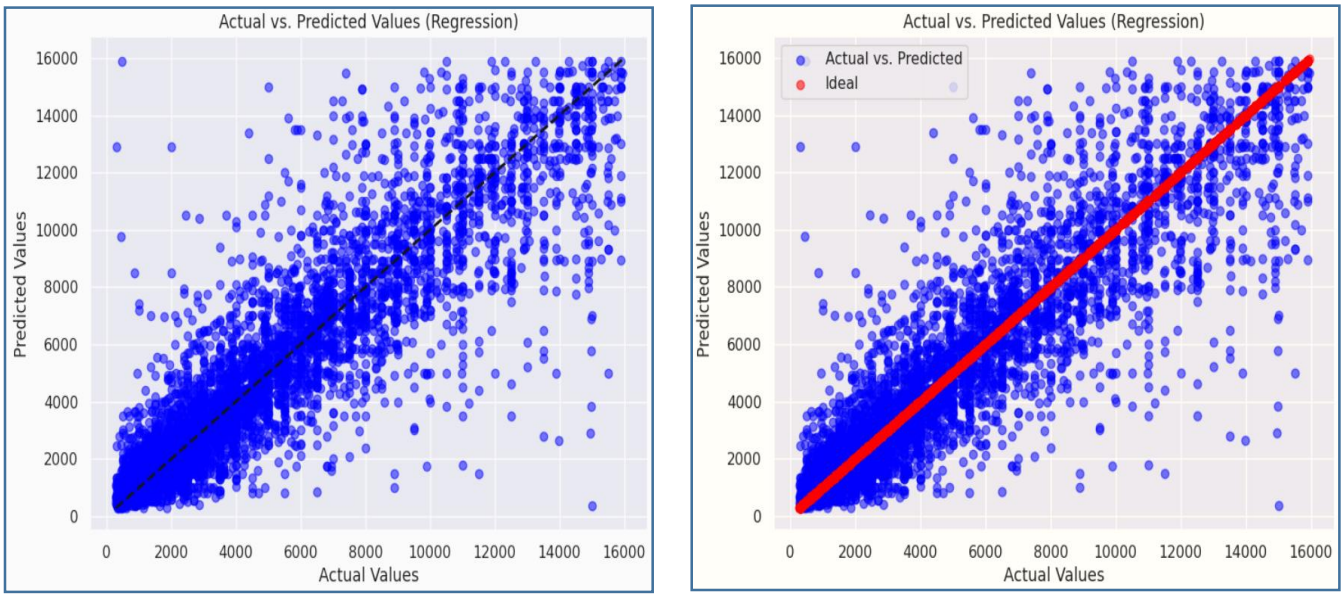

Figure 3: Observed vs Predicted Price – Regression Tree

10.2Residual plot

A residual plot is a graphical tool used in statistics and regression analysis to evaluate the quality of a regression model, specifically to look at the residuals of the model. It is also referred to as a residual analysis plot or a residual vs. fitted values plot. The discrepancies between observed or actual values and predicted values are known as residuals, and they indicate the model's errors. Plotting the residuals on the y-axis and the matching predicted (fitted) values on the x-axis results in a residual plot. A residual plot's objective is to visually examine the residuals for trends or regular departures from randomness.

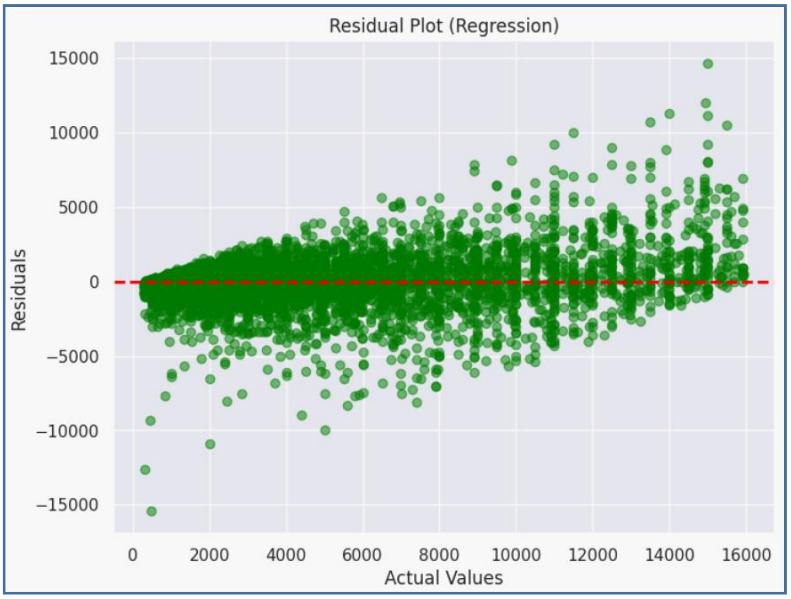

Figure 4: Residual Plot (Regression)

In regression analysis, residual plots are crucial for validating and diagnosing the model. The residuals can be visually inspected to help us spot possible issues with the model and make the necessary corrections to increase accuracy and dependability. A good indication that the regression model fits the data well is if the residuals seem to be randomly distributed around zero.

11. Result:

11.1 Result & discussion:

Communication of a study's findings to readers is facilitated by the inclusion of the Results and Discussion sections, which explain the significance of the findings and how they add to the body of knowledge already in existence. An overview of the research project is provided by these sections, which are frequently accompanied by an introduction, methodology, and conclusion. To fulfill our research, we have used so many terms and condition. We use all of those terms for getting a great result, and we basically use 1) Presentation of Data 2) Accuracy and Objectivity 3) Use of

Visuals 4) Tables and Figures and after using all of this term we can finally get the actual result. From the first to the end we have use all the process gradually and we have also use around three types of models and the best model is “Decision Tree” because it gives the best result among the three models.

Model Name	R2 score	Mean squared error	Root Mean square error	Mean absolute error
Random forest	0.9559710758029846	673746.2754864842	820.8204891975371	533.2398717847389
Linear regression	0.7122655749443627	4390265.816213345	2095.296116593868	1598.0142136235274
Decision tree regressior	0.9661673614633112	516220.04009845643	718.4845440915598	349.9264161498631
KNN	0.872465184292978	1945932.405089663	1394.966811465299	929.6161680078321
Neural Network Regression	0.8039699925719461	2991035.363241713	1729.4610036776526	1191.8910632838395
Gradient Boosting Regression	0.8420636704979133	2409800.177428001	1552.353109774964	1086.62353247784

Performance Measurement Matrices result table

All the methods of Decision Tree that we have used they gives us the best result as we saw from the table before. Each of the model's prediction error rate was significantly lower than the permitted 5% error rate. Upon additional examination, however, it was discovered that the regression tree model's mean error exceeded the multiple regression and linear regression models' mean error rates. The regression tree's accuracy may be higher for every seeds, and its error rates are lower from the remaining seeds.

11.2 Comparative analysis:

Our method has shown an outstanding result compare to other results. As we see that we have work on properly and try to get better result. And we have examined other papers, they have also use so many models and get their result. But our result is better than other results. Our accuracy rate is around 96.6% which is better than other

models. It is noteworthy that not every study endeavor yields revolutionary revelations, and the results may differ. While some research may only make minor improvements, others might have a more profound effect. Research, regardless of size, is an essential procedure for expanding information and tackling issues and problems in a variety of fields.

paper	Model	Accuracy
[11]	ANN	92.38%
[12]	K nearest Neighbor algorithm	85%
[13]	Decision Tree	95%
Our study	Decision Tree	96.6%

Findings something different from research can be important in developing a thorough grasp of a subject, validating conclusions, or pinpointing areas in need of additional study. A methodical and deliberate approach is necessary when comparing research findings. Understanding the research questions, methods, and settings of the studies being compared is crucial when approaching the comparison. Furthermore, researchers should avoid drawing too much conclusions from their work and keep an open mind about the possibility of different results.

12.Conclusion:

Accurately predicting the price of a used car is difficult due to the large number of features and parameters. The collection and preprocessing of data is the first and most important step. After that, a model was developed and defined in order to apply algorithms and produce outcomes. The Decision Tree Algorithm was found to be the best performer after several regression algorithms were applied to the model. Its highest r2 score of 96.6% simply indicated that it produced the most accurate predictions, as shown by the Original vs. Prediction line graph. In addition to having the highest r2 score, Decision Tree also had the lowest Mean Squared Error and Root Mean Squared Values, indicating that its predictions had the fewest errors overall and that the results it produced were therefore extremely accurate.

Reference:

[1] Sadhvik, Kandugula, et al. "Car-Economics: Forecasting Prices in the Pre-Owned Market Using Machine Learning."

[2] Yang, R. R., Chen, S., & Chou, E. (2018). AI blue book: vehicle price prediction using visual features. arXiv preprint arXiv:1803.11227.

[3] Gegic, Enis, et al. "Car price prediction using machine learning techniques." TEM Journal 8.1 (2019): 113.

[4] Gajera, Prashant, Akshay Gondaliya, and Jenish Kavathiya. "Old car price prediction with machine learning." Int. Res. J. Mod. Eng. Technol. Sci 3 (2021): 284-290.

[5] Von Dollen, David, et al. "Quantum-Assisted Feature Selection for Vehicle Price Prediction Modeling." arXiv preprint arXiv:2104.04049 (2021).

[6] AlShared, Abdulla. "Used Cars Price Prediction and Valuation using Data Mining Techniques." (2021).

[7] Meng, Jiajia. "A used car repricing method based on K-Means++ clustering and multiple linear regression [J]." Academic Journal of Business & Management 5.5 (2023): 1-8. [8] Samruddhi, K., & Kumar, R. A. (2020). Used car price prediction using k-nearest neighbor based model. Int. J. Innov. Res. Appl. Sci. Eng.(IJIRASE), 4, 629-632.

[9] Du, Shuyang, Haoli Guo, and Andrew Simpson. "Self-driving car steering angle prediction based on image recognition." arXiv preprint arXiv:1912.05440 (2019).

[10] Kalpana, G., Durga, A. K., Reddy, T. A., & Karuna, G. Predicting the Price of Pre-Owned Cars Using Machine Learning and Data Science.

[11] Gegic, Enis, et al. "Car price prediction using machine learning techniques." TEM Journal 8.1 (2019): 113.

[12] Samruddhi, K., and R. Ashok Kumar. "Used car price prediction using k-nearest neighbor based model." Int. J. Innov. Res. Appl. Sci. Eng.(IJIRASE) 4 (2020): 629-632.

[13] Sharma, Ashutosh Datt, et al. "Predictive analysis of used car prices using machine learning." Int. Res. J. Modernization Eng. Technol. Sci 3 (2021): 674-684.