ABSTRACT

Machine learning isn't just useful for predictive texting or smartphone voice recognition, it's constantly being applied to new industries and new problems. It is the base for multiple things we do online like computer translation & search engines. This world is filled with humongous data which comes with the task of interpreting it. Machine learning helps us build a computer program which increases the presentation of the data to be learned and gives us desired output after understanding the input.

With an increase in demand for automobiles, the production of cars has increased several fold, over 70 million cars were produced in a single year (2016). The market of second hand cars has steadily increased due to this, resulting in the global rise of customers who due to their financial incapability can't buy them. The existing market system decides the prices for these cars randomly and thus loses the cars existing value.

In this research paper we will be comparing different ML algorithm models that can be used to estimate used cars prices and thus will conclude to best working model

After a brief introduction of relevant branches such as AI, and Machine Learning, the paper will then encompass the classified techniques models of Linear Regression which is a linear approach to predicting the relationship between
one dependent and one or more independent variables, followed by Ridge Regression , Lasso Regression .The purpose of this term paper is to create a fully functional model by the usage of examples in the present day scenario as well as any future prospects it might offer.

Fig.1- **Predicting used car prices using ML and AI**

INTRODUCTION

AI has already been touching our lives in ways you might over look for example the as soon as we ask any question to siri it starts using human language processing and speech recognition to perform the task . A simple definition for Artificial intelligence could be, the in which machines are artificially incorporated with human-like intelligence to perform tasks as we do, this intelligence is built using complex algorithms and mathematical functions. In recent years it has made creation of algorithms that can replace the Statistical Techniques possible. AI provides machines with the capability to adapt to environment, reason with given data and then provide solutions.

Often one tend to have a common misconception that machine learning and artificial intelligence are the same cause the boundary of these terms are not clear. Artificial intelligence is the science of getting machines to mimic the behaviour of humans and ML is a subpart of artificial intelligence that focuses on getting machines to make decisions by feeding them data .Machine learning is a subfield of artificial intelligence. With the usage of artificial intelligence, we can analyse the huge amount of information in short time.

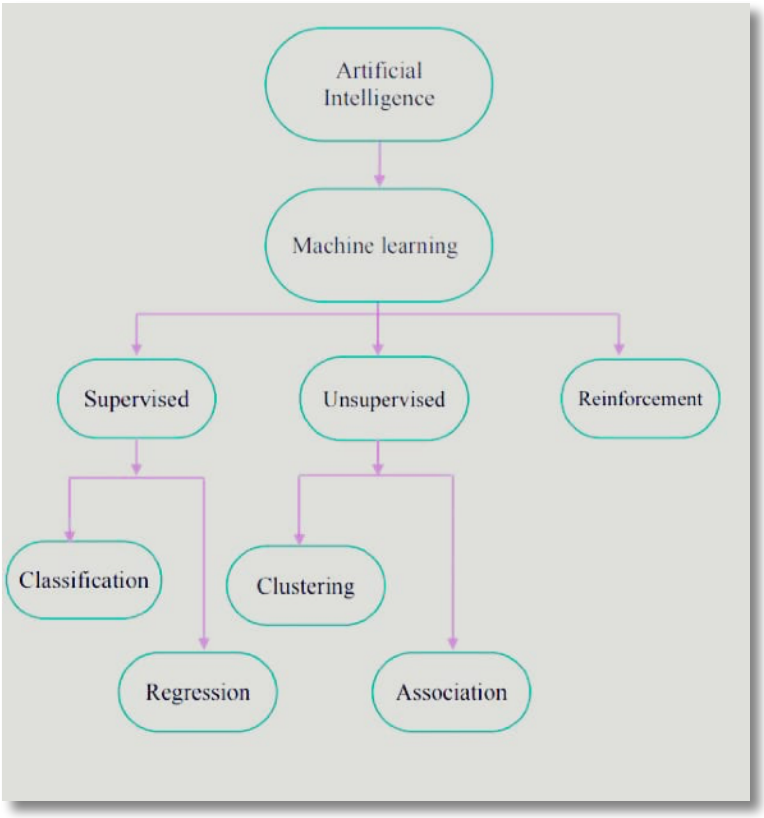

Fig. 2- AI and its subsets

Machine learning , the term is pretty self-explanatory we want to make the machine learn and make decisions based on its knowledge. Due to the humongous data generated not only by people, but also by computers, phones and other devices human brain is incapable to solve a large problem amount of data and complicated relations so it is very much necessary to align and interpret the data and ML method are used in analysing this types of data , it takes data as input interpret it with the help of algorithm models and give the desired output. Machine learning provides a machine with the capability to learn from data and experience through algorithm without being explicitly programmed.

Two types of ML techniques are in use supervised and unsupervised . Supervised learning as the name suggest works under supervision i.e a model which is able to predict output with the help of well labeled data . In unsupervised learning there is no supervision i.e no training will be given to the machine allowing it to act on the data which is not labeled hence machine try to identify patterns and give the output .

In this paper We will compare the performance of distinctive algorithms of machine learning like Linear Regression, Ridge Regression, Lasso Regression, random forest algorithm and choose the best out of it.

When the output variable has real and continuous value it is termed as regression algorithm. Relationship between two or more variables where a change in one variable is associated with change in other is also called regression.Regression Algorithms are used because they provide us with continuous value as an output and not a categorised value because of which it will be achievable to project the actual price a car rather than the range of a automobile.

OBJECTIVE

- In this paper, we conducted a comparative study using multiple linear regression, Ridge Regression , Lasso Regression and and random forest regression build a price model of used car.
- To achieve utmost accuracy.
- The key aim of this paper is to find the finest predictive model for predicting used car price.

METHODOLOGY

Machine learning is using data to answer questions, to explain further machine takes data that they analyse and then characterise through algorithm models for returning output . This is done in two phrase in the system .

- Training phase - Training refers to using our data to inform the creation and fine tuning of a predictive model.This predictive model can then be accustomed cater predictions on previously unseen data and answer those questions. The system is trained using data sets which are then fitted into a model based on the chosen algorithm. Training is a phrase of constructing a model from the training set after this the model can give new output for new inputs , even though for inputs non existing in the training set.

- Testing phrase - in this phrase the system is provided with completely different set of data from the training phrase as input and then it is tested for its working. The accuracy of prediction is checked. Testing is done before choosing the algorithms for further use, different algorithms are compared for their accuracy. The most accurate one for the task is chosen.

PROPOSED SYSTEM

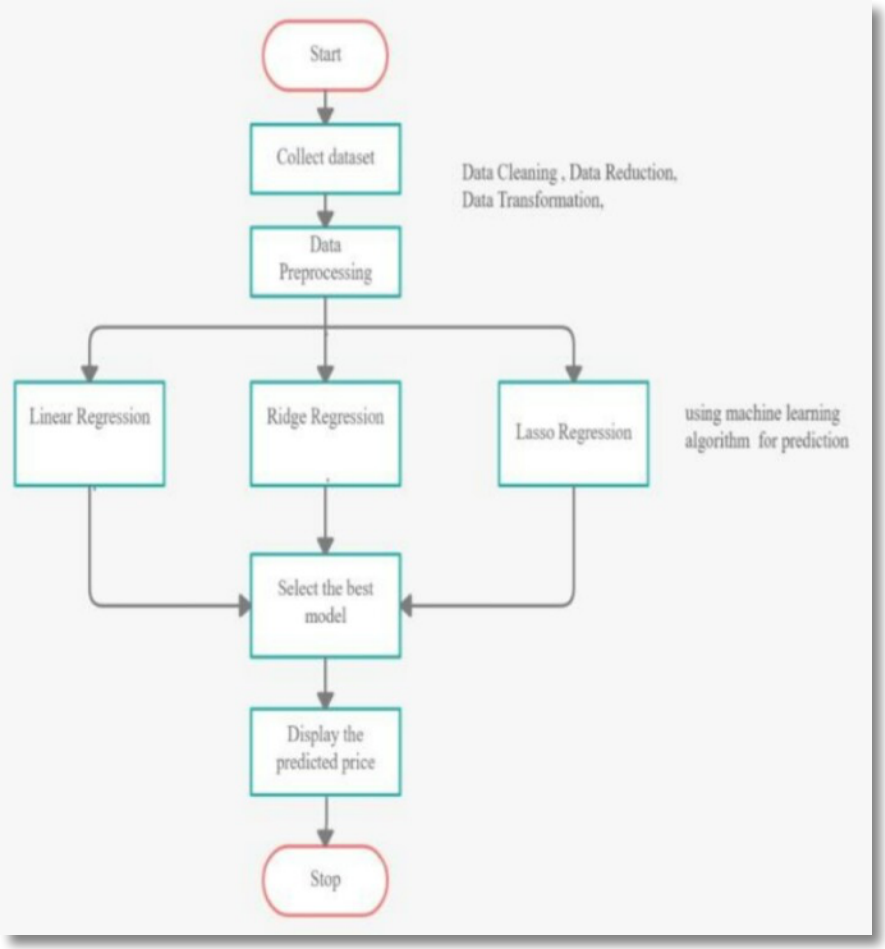

Fig. 3- Proposed system flowchart

As shown within the above figure, the method starts by collecting the dataset. the following step is to try to do Data Preprocessing which includes Data cleaning, Data reduction, Data Transformation. Then, using various machine learning algorithms like linear regression, rigid regression and lasso regression we are going to predict the price of used cars. For this paper, we will be involving algorithms like linear regression, Ridge Regression, and Lasso Regression. The most effective model which predicts the foremost accurate price is chosen. After the selection of the most accurate model, the anticipated price is flaunted to the user per user's inputs. Input can be given by the user through the website to machine learning model for used car price prediction.

Regression based models are proven more efficient and reliable for prediction of continuous variables in supervised learning . For prediction of used cars prices we have to take several continuous variables into consideration.

LINEAR REGRESSION

Linear regression is One the simplest and widely notable machine learning algorithm falling under regression. Linear regression has been taken by machine learning from the field of statistics and it is used as a model to understand the connection between the numerical variables of input and output. In simpler words, it fits two variables into a linear equation which is made with consideration to observed data in attempt to project the relationship between them. Like for example : linear regression model can be used by a modeler to show the relationship between an individuals age and their weight. This is mostly calculated by least square method, in other words we find the line that results in the minimum sum of squared residuals.

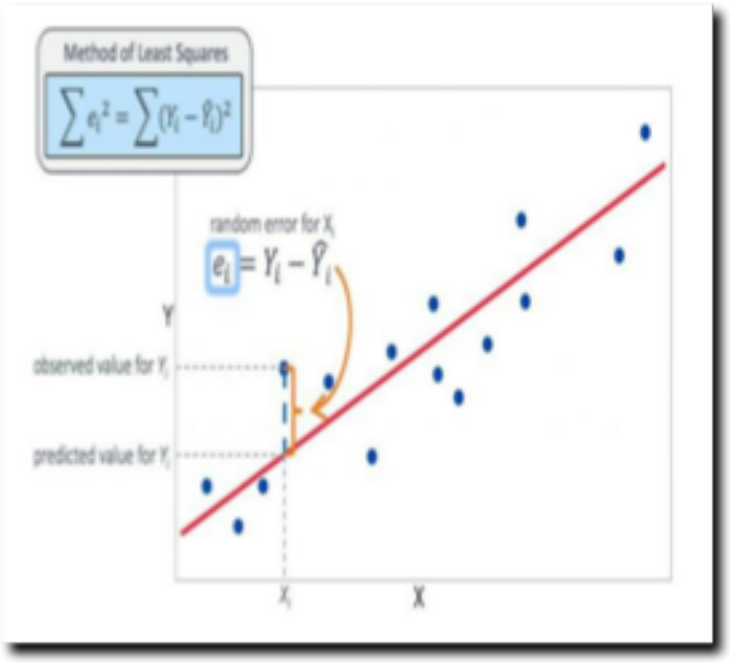

Fig. 4- Linear regression

Mathematically, If the two variables have any relation then This linear model calculates the only output variable (y) from the linear combination of input variables (x). This predicts that 'y' is depends on 'x'.

Linear regression can be classified into types :-

(i) Simple linear regression : As the name suggests when we have only one input variable to calculate the output variable then the method used is known as simple linear regression. Equation (1) gives its mathematical formula.

(ii) Multiple linear regression : whenever there are two or more than two input variables to evaluate the output variable then multiple linear regression becomes operative.(2)

$$y = mx+c \quad (1)$$

where, m is slope of the line, c is y -intercept , y is dependent variable and x is independent variable.

This formula can be further evolved for multiple linear regression.

$$y = m1x1+m2x2+\ldots\ldots+c \quad (2)$$

Where; m1, m2, m3…. = slope, c= y intercept , x1,x2,x3= independent variables and y= dependent variables.

RIDGE REGRESSION

Ridge regression aids in creating a simple model with noteworthy explanatory predictive power that explain data within minimal number of parameters . This becomes active when a data set has multicollinearity (correlations between predictor variables) or when the number of predictor variables (an independent variable put to use in regression analyses) in a set exceeds the number of observations. Basically regularised extension of linear regression is know as ridge regression . It is called also L2 regularization.

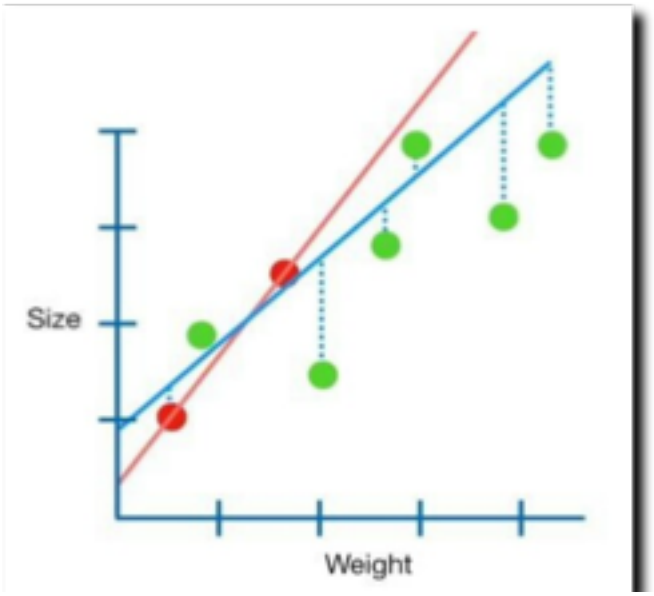

Fig. 5 - Ridge regression

The basic idea behind Ridge regression is to look for a new line on the graphical representation that doesn't align with the training data as well , in other words we introduce a small amount of bias into how the new line is fit to the data but in return for that small amount of bias we get a significant drop in variance in other words by starting with a slightly worse fit Ridge regression can provide better long-term predictions.Ridge regression advantage is to mainly avoid overfitting. Overfitting in general occurs when the trained model performs well during training and performs very poorly during testing.

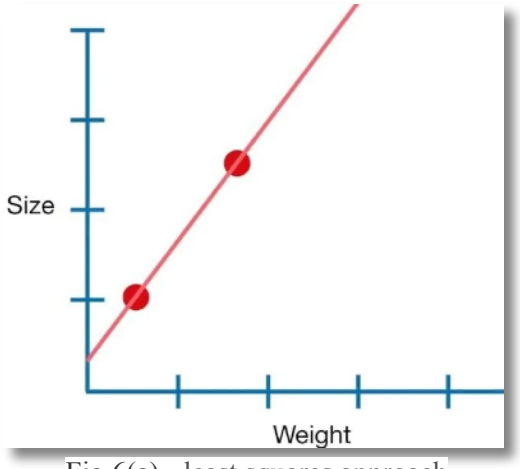

Fig 6(a) - least squares approach

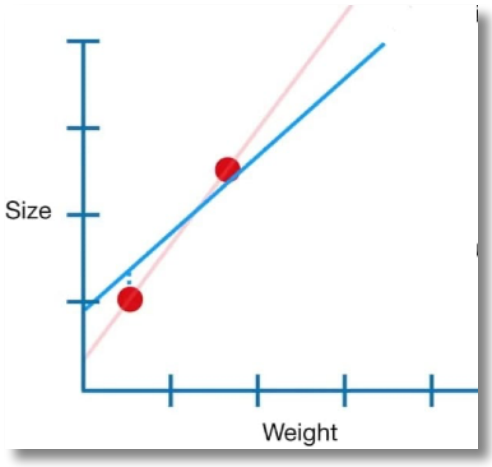

Fig 6(b) - Ridge regression approach

Fig 6(a)- Shows that when least-squares determines values for the parameters in the equation it minimises the sum of the squared residuals .

Size = y-axis intercept + slope X **weight**

Fig 6(b). Shows when Ridge regression determines the values for the parameters in the equation it minimises the sum of the squared residuals plus alpha times the slope squared

Size = y-axis intercept + slope X **weight** + **α** X **slope**^{2}

Slope^2 part of the equation adds a penalty to the traditional least squares method and alpha (reduces variance of the estimates) determines how severe that penalty will be.

From the above sample graph we can conclude that if we want to minimise the sum of the squared residuals plus the ridge regression penalty we should choose the ridge regression line over the least squares line. Without the small amount of bias that the penalty creates the least squares fit has a large amount of variance in contrast the ridge regression line which has a minute amount of bias due to the penalty has less variance.

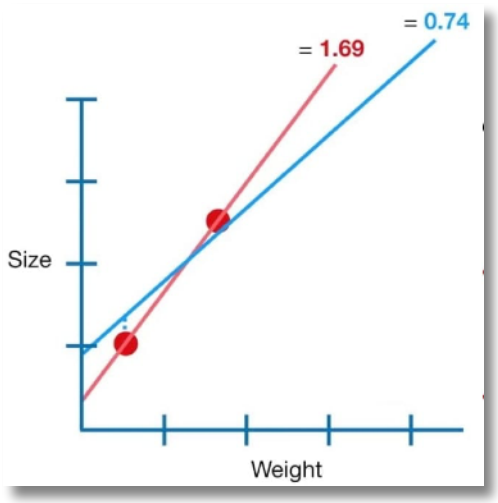

Fig 7 - Experimental data results

LASSO REGRESSION

The abbreviation 'LASSO' put together the words Least Absolute Shrinkage and Selection Operator. Like ridge regression, lasso regression is also a regularisation technique. Lasso regression is very much similar to the ridge regression it also works by introducing a bias term in the equation but instead of squaring the slope the absolute value of the slope is added as a penalty term. The big difference between Ridge and lasso regression is that Ridge regression can only shrink a slope asymptotically close to zero while lasso regression can shrink the slope all the way to 0 .

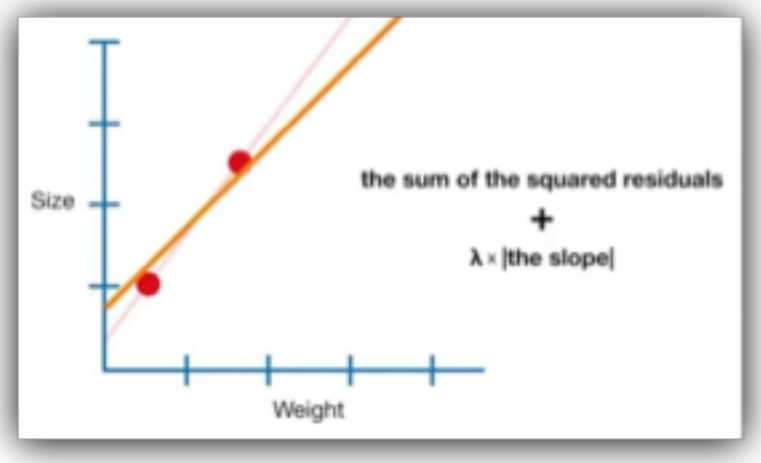

Fig 8 - Lasso regression

when lasso regression determines the values for the parameters in the equation it minimises the sum of the squared residuals plus alpha times the absolute value.

Size = y-axis intercept + slope X **weight** + **α** X **|the slope|**

Since lasso regression can exclude useless variables from equations it is a little better than Ridge regression at reducing the variance and models that contain a lot of useless variables in contrast ,Ridge regression performs better when most variables are useful. Lasso regression makes the final equation simpler and easier to interpret. This model uses shrinkage. The data values are shrunk towards a central or median point as the mean in shrinkage.

FUTURE SCOPE

Price predicting real time data can be provide through various websites when they are combined with machine learning models in near future. To enhance the accuracy percentage of these machine learning models we can introduce considerable amount of historical data of car prices. For the betterment of used car market websites should build bots and programs that focuses on user interaction . Price prediction of used cars is quite an interesting and well known problem. As per information that was gotten from the NTA(National Transport Authority) a 234% increase in the no of cars registered was seen ,within a time spam of 10 years (2003-2013) the number of registered cars have reached 160,701 in 2013 from 68,525 registered cars in 2003. number of cars will increase in future. As it is very likely that this tend will continue . Specially now, cause of the economic condition has worsened because of covid so people are more likely to buy more second hand cars resulting in an increment of its sales ,additional significance is adde to the problem of the car price prediction cause of this . Although, this research paper has achieved desired results our aim.

Fig.9- future scope

INTERNATIONAL & NATIONAL STATUS OF WORK

In today's era , we have big data available on various platforms, everyone is either just surfing the net or are making some kind of online transactions which is generating more and more amount of data with time and now that we know data plays a key role in machine learning thus with increasing amount of data the machine learning opportunities will also increase . And following the general trends, it can be noted that 'Supervised Learning' (training algorithms with manually labeled datasets, instead of letting the network find a hidden pattern) will be the blooming area in coming years, for both India and overseas.

Used Car Market Report Scope

Report Attribute	Details
Market size value in 2020	USD 1,402.0 billion
Revenue forecast in 2027	USD 2,150.6 billion
Growth Rate	CAGR of 5.5% from 2020 to 2027
Base year for estimation	2019
Historical data	2016 - 2018
Forecast period	2020 - 2027
Quantitative units	Revenue in USD billion, shipment in million units, and CAGR from 2020 to 2027
Report coverage	Revenue forecast, company ranking, competitive landscape, growth factors, and trends
Segment Scope	Vehicle type, vendor type, size, sales channel, fuel type, region
Region scope	North America; Europe; Asia Pacific; South America; MEA
Country scope	The U.S.; Canada; Germany; The U.K.; France; Spain; China; Japan; India; Brazil
Key companies profiled	Alibaba.com; Asbury Automotive Group; AutoNation Inc.; CarMax Business Services, LLC; Cox Automotive; eBay Inc.; Group 1 Automotive Inc.; Hendrick Automotive Group; LITHIA Motor Inc.; Scout24 AG; TrueCar, Inc.
Customization scope	Free report customization (equivalent up to 8 analysts working days) with purchase. Addition or alteration to country, regional & segment scope.
Pricing and purchase options	Avail customized purchase options to meet your exact research needs. Explore purchase options

Fig.10- Global scope

CONCLUSION

The increased prices of recent cars and also the financial incapability of the shoppers to buy them. Sales of used cars are at spring peak worldwide. Therefore, there's an urgent need for a second hand Car Price Prediction system which effectively determines the worthiness of the car employing a kind of features. It is challenging to get accurate prediction due to great number of attributes involved. The proposed algorithms will help in predicting the accurate worth of used car. In this paper we looked at 3 distinctive machine learning algorithms :- Linear Regression, Lasso Regression and Ridge Regression. By the end we concluded As more data is gathered, the model can be improved over time and new predictive models deployed. 2