

SALES FORECAST OF MANUFACTURING COMPANIES USING MACHINE LEARNING NAVIGATING THE PANDEMIC LIKE COVID-19

Submitted in partial fulfillment of the requirements
for the award of Bachelor of Engineering degree
in Computer Science and Engineering

By

Venkata Manas Cheruku (381100634)

Kakumanu Eswar (38110222)



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SCHOOL OF COMPUTING

**SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited with Grade “A” by NAAC**

**JEPPIAAR NAGAR, RAJIV GANDHI SALAI,
CHENNAI - 600 119**

MARCH – 2022



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY



(DEEMED TO BE UNIVERSITY)

Accredited with “A” grade by NAAC

Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai - 600119

www.sathyabama.ac.in

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work **Venkata Manas Cheruku (381100634)** and **Kakumanu Eswar (38110222)** who carried out the project entitled “**SALES FORECAST OF MANUFACTURING COMPANIES USING MACHINE LEARNING NAVIGATING THE PANDEMIC LIKE COVID-19**” under my supervision from November 2021 to March 2022.

Internal Guide

Dr.S.Murugan M.E., Ph.D.

Head of the Department

Dr. S. Vigneshwari M.E., Ph.D., and Dr. L. Lakshmanan M.E., Ph.D.,

Submitted for Viva voce Examination held on _____

Internal Examiner

External Examiner

DECLARATION

I **Venkata Manas Cheruku (381100634)** and **Kakumanu Eswar (38110222)** hereby declare that the Project Report entitled “**SALES FORECAST OF MANUFACTURING COMPANIES USING MACHINE LEARNING NAVIGATING THE PANDEMIC LIKE COVID-19**” done by us under the guidance of **Dr S.Murugan M.E., Ph.D.** is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering degree in 2018-2022.

DATE:

PLACE:

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management of SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T.Sasikala M.E.,Ph.D., Dean**, School of Computing **Dr.S.Vigneshwari M.E., Ph.D.** and **Dr.L.Lakshmanan M.E., Ph.D.** , Heads of the Department of Computer Science and Engineering for providing us necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and a deep sense of gratitude to my Project Guide **Dr. S. Murugan M.E., Ph.D.**, for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

TABLE OF CONTENTS

Chapter No.	TITLE	Page No.
	ABSTRACT	8
	LIST OF FIGURES	vii
1	INTRODUCTION	9
2	LITERATURE SURVEY	10
	2.1. Car Sales Prediction Using Machine Learning Algorithms <i>Abstract</i>	10
	2.2. Predicting the Price of Used Cars using Machine Learning TechniquesAbstract	18
	2.3. Predicting the Price of Second-hand Cars using Artificial Neural Networks	25
	2.4 USED CAR PRICE PREDICTION	30

3	METHODOLOGY	36
	3.1. EXISTING SYSTEM	36
	3.2. PROPOSED SYSTEM	37
	3.3. SYSTEM ARCHITECTURE	38
	3.4. WORKING OF DECISION TREE	39
4	RESULTS AND DISCUSSION	54
	4.1. MACHINE LEARNING	
5	CONCLUSION	64
	5.1. CONCLUSION	
6	CONCLUSION	67
	6.1. CONCLUSION	
	REFERENCES	69
	APPENDICES	69
	A. SOURCE CODE	69
	B. SCREENSHOTS	89
	C. PLAGIARISM REPORT	93

94

LIST OF FIGURES

Figure No.	Figure Name	Page No.
3.1	Block Diagram	37
3.2	Flow Diagram	38
3.3	Real Life Analogy	47
3.4	Bagging	49
3.5	Bagging Ensemble Method	49
3.6	Model	53
4.1	Clustering	61
4.2	ML Overview	62
4.3	Training and Testing	
4.4	Validation Test	
4.5		

ABSTARCT:

The paper is concerned with statistical models to forecast resale prices of used cars. An empirical study is performed to explore how different degrees of freedom in the modeling process contribute toward forecast accuracy. First, a comparative analysis of alternative prediction methods evidences that random forest regression is particularly effective in resale price forecasting. It is also shown that the use of linear regression, the prevailing method in previous work, should be avoided. Second, empirical results evidence the presence of heterogeneity in resale price forecasting and identify methods that can automatically overcome its detrimental effect on forecast accuracy.

Finally, the study confirms that the sellers of used cars possess informational advantages over market research agencies, which enable them to forecast resale prices more accurately. This implies that sellers have an incentive to invest into an in-house forecasting solution, instead of basing pricing decisions on externally generated residual value estimates.

Chapter 1

INTRODUCTION:

In many developed countries, it is common to lease a car rather than buying it outright. A lease is a binding contract between a buyer and a seller (or a third party – usually a bank, insurance firm or other financial institutions) in which the buyer must pay fixed installments for a pre-defined number of months/years to the seller/financier. After the lease period is over, the buyer has the possibility to buy the car at its residual value, i.e. its expected resale value. Thus, it is of commercial interest to seller/financiers to be able to predict the salvage value (residual value) of cars with accuracy. If the residual value is under-estimated by the seller/financier at the beginning, the installments will be higher for the clients who will certainly then opt for another seller/financier. If the residual value is over-estimated, the installments will be lower for the clients but then the seller/financier may have much difficulty at selling these high-priced used cars at this over-estimated residual value. Predicting the resale value of a car is not a simple task. It is trite knowledge that the value of used cars depends on a number of factors. The most important ones are usually the age of the car, its make (and model), the origin of the car (the original country of the manufacturer), its mileage (the number of kilometers it has run) and its horsepower. Due to rising fuel prices, fuel economy is also of prime importance. Unfortunately, in practice, most people do not know exactly how much fuel their car consumes for each km driven. Other factors such as the type of fuel it uses, the interior style, the braking system, acceleration, the volume of its cylinders (measured in cc), safety index, its size, number of doors, paint colour, weight of the car, consumer reviews, prestigious awards won by the car manufacturer, its physical state, whether it is a sports car, whether it has cruise control, whether it is automatic or manual transmission, whether it belonged to an individual or a company and other options such as air conditioner, sound system, power steering, cosmic wheels, GPS navigator all may influence the price as well. Some special factors which buyers attach importance in Mauritius is the local of previous owners, whether the car had been involved in serious accidents and whether it is a lady-driven car. The look and feel of the car certainly contributes a lot to the price. As we can see, the price depends on a large number of factors. Unfortunately, information about all these factors are not always available and the buyer must make the decision to purchase at a certain price based on few factors only. In this work, we have considered only a small subset of the factors mentioned above. More details are provided in Section III. This paper is organised as follows. In the next section, a review of related work is provided. Section III describes the methodology while in section IV, we describe, evaluate and

compare different machine learning techniques to predict the price of used cars. Finally, we end the paper with a conclusion with some pointers towards future work.

Chapter 2

Literature review

2.1 Car Sales Prediction Using Machine Learning Algorithms

Abstract:

Sales prediction is the current numero trend in which all the business companies thrive and it also aids the organization or concern in determining the future goals for it and its plan and procedure to achieve it. The data about car sales are derived from various sources .sales of cars does not contain any independent variable since various factors such as horse power; model, width, fuel type, height, price, city-mileage, highway-mileage and manufacturer are the various features that influence the sales. In car sales prediction we first implement the methodology of analytic hierarchy process in order to get varied idea about how well the various criteria's in our dataset works and after this we apply the machine learning algorithms such as Linear regression, Random tree to get the best clusters and we process them in to random forest to get best accurate feature out of it. which is ultimately followed by Technique for Order of preference by similarity to ideal solution (TOPSIS) an tool which helps the researcher to arrive at a verdict when he/she faces the one or more pattern selection problem and the final resultant derived from all these methods gives the fittest feature which influences the customer in purchasing the car which indirectly gives the company or the research market a result in predicting the future sales for cars. Hence this paper not only provides its users with some stats, it also serves an guiding guardian by providing accurate results for purchasing a car.

INTRODUCTION

Every human being has got his/her business in this world .so the term business comes with certain tagged words such as sales, profit and loss. A concerns success is determined by its sales and the performance of its product in the industry is also identified by its sales. Hence in order to improve the standard of the firm /concern the strategies, the techniques the methodologies are built .in this process of development forecasting of certain key terms such as profit, loss, return of interest may pave the way for the successful venture of the concern yet when we look deep in to the details all

these key terms are conveniently related to the term called sales .sales prediction is one of the master trades of business which may open the gateways for obtaining knowledge about the existing market trends and the ways to conquer the market .planning is the first step in every activity we perform and hence knowing what lies ahead in terms of sales hugely aids the concern/organization in this planning process . sales prediction can be a massive support in order to perform tracking, cashflow and purchasing .sales forecasting provides the business minds an idea of about how much to buy and how much not to buy ,what are the risk can be taken in terms of revenue ,how to plan budget ,market trends, introduction of new products according to the organizations capability and ability ,what changes may happen if the plan fails are the areas where it helps to develop our ideas ,let us shift our focus from the generic business term to its detailed chunk of streams such as education,medical,automobile and lots of other streams. Once great interest in recent days lies in automobile industry. The first boon to the automobile industry came only after the industrialization renaissance period. Now the automobile industry creates, packs, runs, moves the world on its wheels and also allows other fields to flourish parallel along with them .yet it also has its own defaults since it is not cost friendly ,any single default can affect the whole system and it cause a huge failure in the market for its brand ,there is a huge amount of revenue being spend and huge amount of chunks being produced and exploited on daily basis hence sales forecasting in this industry can lead to the organized pattern of selling,buying,producing goods and even the taxes to be implemented also comes in to role .forecasting these sales in automobile industries can be performed with various and variety of technologies and one among them is the machine learning technique. That may help in classifying the prediction of an automobile for a say lets us take it as car the yearly sales of it if know before then it will provide the manufacturer the huge boost in designing it, getting spare parts, getting key parts and reducing the waste products and tracking its revenue model its generation and various other activity. The classifiers used such as logistic regression, decision tree and random forest provides us with accurate prediction results.

Literature Review

Machine learning models and bankruptcy prediction is a paper work which talks about the improvement that takes place in academics industry with the aid of machine learning algorithms in predicting bankruptcy. The data is derived from integrated resource of Salomon center database which contains the details about the North American firms from the period between 1985 to 2013 . This paper implements the usage of algorithms such as bagging, boosting, random forest and support vector machine for predicting bankruptcy even before the event occurs and a greater span

of comparative study takes place with the performance of these results with the results of logistic regression and neural networks. Original Altman's Z-score variables are used as predictive variables with addition of extra variables such as the operating margin, sales, growth measures related to assets, change in return-on-equity, change in price-to-book, and number of employees based on carton and Hofer(2006). And a comparison is made between the models and these variables, the machine learning techniques and the algorithm with most accuracy is determined. Handling class imbalance in customer churn prediction by j.Burez and D. van den poel suggests the customer the various ways to handle class imbalance in churn prediction. AUC and lift are the evaluation metrics with which the sampling methods are interrogated. The modeling techniques such as weighted random forest, gradient boosting are compared with other techniques. The better evaluation metrics and the best modelling techniques are found out with the help of each techniques accuracy and from past studies. Calling communities analysis and identification using machine learning techniques is the work that determines the worth of a particular customer with respect to his/her general pattern trait of the community that he/she hails from. The customer calling impressions can be told beforehand by making use of a classifier model and cluster analysis for detail selection. The attributes such as accuracy and computational performance are taken in to consideration for comparison of various machine learning techniques. Customer churn prediction using improved balanced random forests is the paper that explains the real time working model that had been used in china. Improved balanced random forest is the hybrid version of balanced random forest and weighted random forest, two interval variables had been introduced such as e and f where e is the middle point and f is the length of interval .Random distribution of these classes are maintained with the help of these variables .Hence it produces more accurate results than its other counterparts. A sampling based sentiment mining approach for e-commerce applications paper puts the limelight on how the customers are being influenced by the Online reviews which is a part of marketing strategy of the e-commerce platforms. Hence this issue is attempted with the help of mining techniques. The two sampling methods are also used for classification of imbalanced data. A modified support vector machine based ensemble algorithm is the methodology used by the researches to identify the performance prediction [10]. On the differential benchmarking of promotional efficiency with machine learning modelling (II): Practical applications presents two different databases of different categories such as non-seasonal and heavy seasonal and models are analyzed here .The detailed performance of four famous machine learning techniques that has huge complexity is been dissected in this work. Certain features of various machine learning algorithms do not perform well because of these databases. In order to gain more accurate dissection results and feature extraction there is a need to implement certain correct procedures that may influence the specificity of the behavior of certain

categories and product ranges. Linguistic features for review helpfulness prediction by Srikumar Krishnamurthy analyses what makes an Online review with the help of a predictive model .This model follows the methodology of extracting linguistic category features such as adjective feature, state verb feature and action verb features it also takes in to account the readability related features for prediction. Hence the hybrid set of features that are obtained after the analysis on two real-life review datasets gives the researches the best accuracy rate of all time. Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data paper puts the focus on how this work can contribute to the real estate industry that may cause an adverse effect on the US housing market.an eight step methodologies are used for the dataset of 5359 townhouses in Fairfax County, Virginia. The dataset has been segregated in to training, validation and testing set then training parameters such as C4.5, RIPPER, Naïve Bayesian, and Adaboost are set and the model is trained and evaluated using training set and validation set respectively and this process is iterated until it gains an optimal error in training, validation and testing.Finally these results are compared to gain the optimal accuracy results. Explaining machine learning models in sales prediction is a generic manuscript that discusses about the recent trends of predictive models, real time scenarios in order to gain a deep insight about buyers and seller's interaction and the forecasting of sales. Early churn prediction with personalized targeting in mobile social games is a manuscript that explains Customer churn .churn is defined by the act of a customer leaving a product for good. This churn are reduced to a greater extent by following the procedure of mapping the feature with the interest of the customer and pushing the notifications in order to drag back the customer in to the game .this manuscript implements the methodologies such as logistic regression for the simple object linear model ,decision trees for extracting redundancy from features random forest to be used in various situations .Naive Bayes for generating the models and gradient boosting for its popularity. Distributed customer behavior prediction using multiplex data: A collaborative MK-SVM approach is a paperwork that explains the challenges that are encountered by the traditional method in predicting customer behavior. It contains a huge datasets which are categorized as static, symbolic sequential, textual data and time series in its database collaborative multiple kernel support vector machine (C-MK-SVM) is the new technique that is used for distributed customer behavior prediction with the aid of multiplex data. Various sub models in this technique is used for global optimization. The results obtained through Computation tell the researches that it is best suited for customer behavior prediction performance and for its maximum computational speed. Customer churn prediction using improved balanced random forest is also one of the works on churn prediction.it undergoes the disadvantage of imbalanced in the data distribution. This improved random balanced forest also uses some other sampling techniques with it. IBRF's features are mis

classified minority class with higher penalties are iteratively learned by altering the class distributions. It works with the real time data such as bank customer database. When compared with other methodologies such as artificial neural networks, decision trees, and class-weighted core support vector machines (CWC-SVM) IBRF has more accurate prediction features. The process /model of the system is kicked off to action by collecting the data for car sales prediction from renowned data repositories and available databases around the world and then these data are preprocessed, then all these data sets are selected for training and the best dataset are classified and these classifiers are further trained using various methodology and the results are predicted and its performance are evaluated and finally the results are being displayed.

IMPLEMENTATION

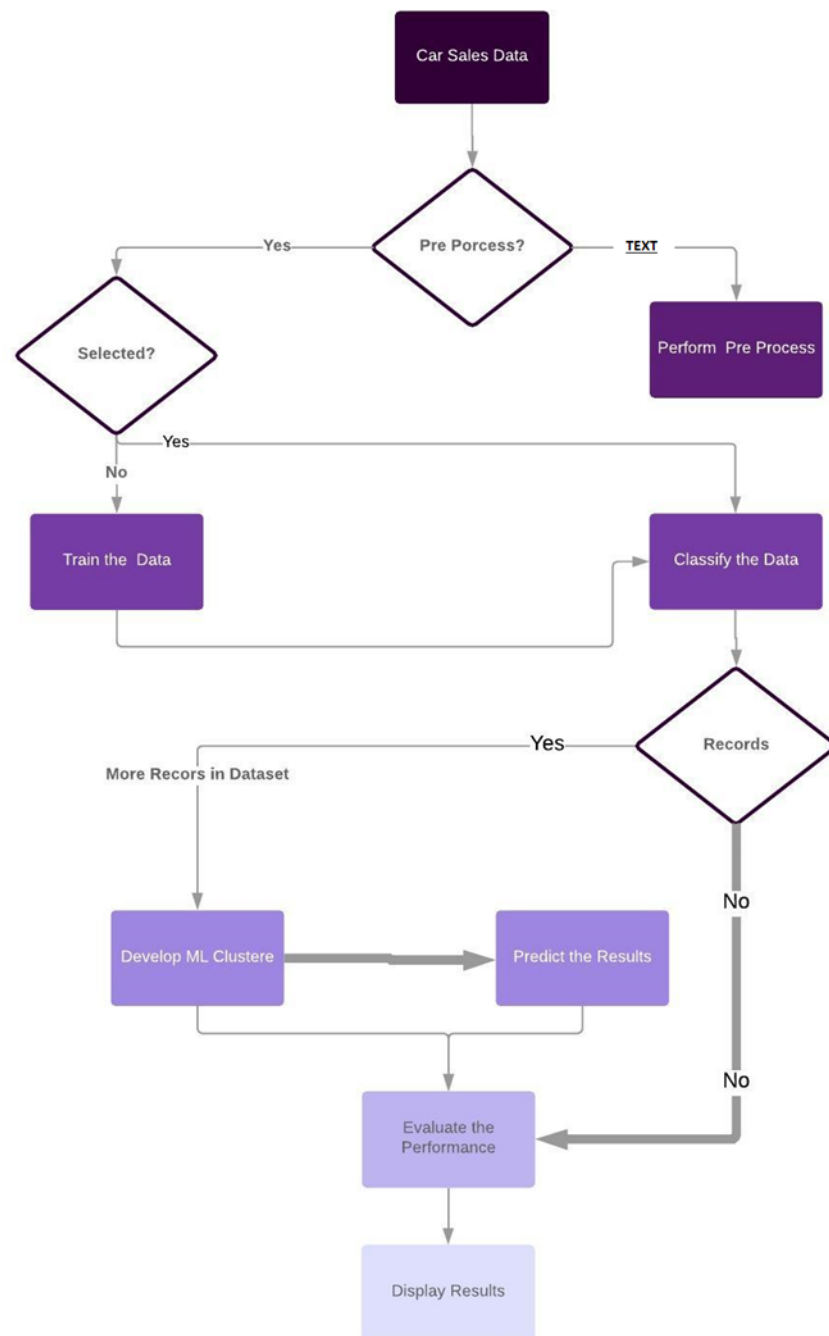
AHP-The Analytic hierarchy process is a Boon to the data industry since it helps people in taking the complex decisions .any decision depends on at what point of time the decision is taken and how well it helps in growth of the product. Hence this AHP helps in taking not just one equitable decision but it stocks up the customer with the most desirable decision solution that can help them in achieving their objective to their problem. It can be implemented in various sectors where we can expect a huge amount of data .For example in automobile industry people can have a problem related to purchasing an automobile in this case AHP comes in to action where out of various types of automobiles it gives the customer a detailed report of desirable solutions. When we look deep down in to AHP process certain steps needs to be followed. We take the Automobile dataset and the attributes of the dataset are given a grading based on the weights of those attributes and in the next process these weighted attributes are been assessed based on its values .The points are pitted against each other in pairs and it is checked whether they can achieve its objective .The number of pairs to be drawn against each other are first entered and then these pairs are pitted against AHP priorities and each attributes are given parallel values in the integer range of one to nine where one represents significant, three represents moderate significance ,five represents strong significance, seven represents very strong significance ,nine represents .in our automobile dataset eight attributes such as city-mileage, highway-mileage, model name,height,width,price,fuel type,horsepower in which price serves with the value of high significance and the difference between each of these eight values are pitted against each other to gain the consistency rate .if that value is less than ten percent it is said to be of with good consistency After the obtaining the suggestion for the most likely decision

like price should have the highest priority the machine learning algorithms are applied to this dataset to obtain a better predictive mode.

Linear Regression:

The interaction that happens between any two fixed components which are also coined as variables and the method used to define or calculate the weightage of their bond is called regression analysis. Their main focus is to make the novice understand the process that happens when there is a drastic change and how these changes happen due to the modification of predictors and the adverse effect it has on criterion variable. At the preliminary level, the data generation serves as a key in order to understand this model. There are always three vital parts needed: the criterion variable, the predictors, and the disguised parameters. We also categorize them into two: simple and multiple (more than one variable). The tools or the metrics to measure the linear variable varies within these two types: an unstopable scale and categorical one. To put in precise terms, correlation is what the linear regression is made of, whereas the difference lies in their ability to distinct things.

Car Sales Prediction Architecture



Random tree:

A random tree is a tree/arborescence which is formed by a random process and can deal with both classification and regression problems. The classifier takes the input feature vector, classifies it with every tree in the forest, and outputs the class label that received the majority of votes. For regression,

the classifier response is the average of the responses over all the trees in the forest. All the trees are trained with the same parameters but on training sets. In random trees, there is no need for an accuracy estimation procedure, such as cross – validation or, a separate test set to get an estimate of the training error.

Random Forest:

This algorithm can be used with ease since it is a cluster of decision trees put together to form a forest so it can be put in to purpose in both classification and regression problems which occur in the field of machine learning. It searches in which part of an attribute set contains the best feature so that it can randomly split them in to a node since there are varsity of attributes they will load us with better performances .and the difference here is only randomly selected attribute set is taken in to consideration for spiting and can be achieved by benchmarking a threshold value for every particular attribute. This methodology also prunes out the most inappropriate tress in the forest to minimize its impurity. During prediction one can easily identify the Significant importance of each feature and a vast difference is also being detected between decision tree and random forest .in order to gain an absolute magnificent performance one needs to tune its hyper parameter hence in this methodology they are max_features, prediction rate, random state and the final parameter for cross validating the random forest. They may be so easy to train but while executing they react in a very slow manner.

TECHNIQUE FOR ORDER OF PREFERENCE BY SIMILARITY TO IDEAL SOLUTION (TOPSIS):

In car sales prediction to give the customer a knowledge about which car can be rated as the best we have certain points of reference where we rate them according to the ranks and present them before the customer that is fuzzy logic. In order to gain the ranks of the car we attribute those points of references with the weighted values. We had already given the numeric preference to the values in AHP method itself and we also follow this here. Mainly we take six attributes such as width, height, horsepower, city, highway and price. We perform vector normalization using formulas such as $E(kl) = E(kl) / \text{Square root of } E^2(kl)$ where summation is from $l=1$ to n . The square root part of this formula is calculated by adding up the square values of a particular column and taking the square root of final results .now the normalized value of each cell is being gained by dividing the performance values with square rooted values and those performance values are replaced by these normalized values hence we get a normalized matrix.

RESULT ANALYSIS

From these studies we infer that price is the attribute that makes an impact on predicting the car sales values, we also rank Chevrolet brand or make of the car as the leading sales in the car industry with the rank of number one which is followed by Honda afterwards. This Rank is being proposed after we perform the “TECHNIQUE FOR ORDER OF PREFERENCE BY SIMILARITY TO IDEAL SOLUTION“. Hence with the help of Analytic hierarchy process, several machine learning algorithms such as linear regression, random forest we have inferred results that are so specific and accurate. The following picture depicts the percentages of the price attribute of each make.

2.2. Predicting the Price of Used Cars using Machine Learning Techniques

Abstract

In this paper, we investigate the application of supervised machine learning techniques to predict the price of used cars in Mauritius. The predictions are based on historical data collected from daily newspapers. Different techniques like multiple linear regression analysis, k-nearest neighbours, naïve bayes and decision trees have been used to make the predictions. The predictions are then evaluated and compared in order to find those which provide the best performances. A seemingly easy problem turned out to be indeed very difficult to resolve with high accuracy. All the four methods provided comparable performance. In the future, we intend to use more sophisticated algorithms to make the predictions.

INTRODUCTION

Predicting the price of used cars is both an important and interesting problem. According to data obtained from the National Transport Authority, the number of cars registered between 2003 and

2013 has witnessed a spectacular increase of 234%. From 68, 524 cars registered in 2003, this number has now reached 160, 701. With difficult economic conditions, it is likely that sales of second-hand imported (reconditioned) cars and used cars will increase. It is reported in that the sales of new cars has registered a decrease of 8% in 2013. In many developed countries, it is common to lease a car rather than buying it outright. A lease is a binding contract between a buyer and a seller (or a third party – usually a bank, insurance firm or other financial institutions) in which the buyer must pay fixed instalments for a pre-defined number of months/years to the seller/financier. After the lease period is over, the buyer has the possibility to buy the car at its residual value, i.e. its expected resale value. Thus, it is of commercial interest to seller/financiers to be able to predict the salvage value (residual value) of cars with accuracy. If the residual value is under-estimated by the seller/financier at the beginning, the instalments will be higher for the clients who will certainly then opt for another seller/financier. If the residual value is over-estimated, the instalments will be lower for the clients but then the seller/financier may have much difficulty at selling these high-priced used cars at this over-estimated residual value. Thus, we can see that estimating the price of used cars is of very high commercial importance as well. Manufacturers' from Germany made a loss of 1 billion Euros in their USA market because of miscalculating the residual value of leased cars. Most individuals in Mauritius who buy new cars are also very apprehensive about the resale value of their cars after certain number of years when they will possibly sell it in the used cars market. Predicting the resale value of a car is not a simple task. It is trite knowledge that the value of used cars depends on a number of factors. The most important ones are usually the age of the car, its make (and model), the origin of the car (the original country of the manufacturer), its mileage (the number of kilometers it has run) and its horsepower. Due to rising fuel prices, fuel economy is also of prime importance.

Unfortunately, in practice, most people do not know exactly how much fuel their car consumes for each km driven. Other factors such as the type of fuel it uses, the interior style, the braking system, acceleration, the volume of its cylinders (measured in cc), safety index, its size, number of doors, paint colour, weight of the car, consumer reviews, prestigious awards won by the car manufacturer, its physical state, whether it is a sports car, whether it has cruise control, whether it is automatic or manual transmission, whether it belonged to an individual or a company and other options such as air conditioner, sound system, power steering, cosmic wheels, GPS navigator all may influence the price as well. Some special factors which buyers attach importance in Mauritius is the local of previous owners, whether the car had been involved in serious accidents and whether it is a lady-driven car. The look and feel of the car certainly contributes a lot to the price.

As we can see, the price depends on a large number of factors. Unfortunately, information about all these factors are not always available and the buyer must make the decision to purchase at a certain price based on few factors only. In this work, we have considered only a small subset of the factors mentioned above. More details are provided in Section III. This paper is organised as follows. In the next section, a review of related work is provided. Section III describes the methodology while in section IV, we describe, evaluate and compare different machine learning techniques to predict the price of used cars. Finally, we end the paper with a conclusion with some pointers towards future work.

Related Work:

Surprisingly, work on estimated the price of used cars is very recent but also very sparse. In her MSc thesis, liztiani showed that the regression mode build using support vector machines (SVM) can estimate the residual price of leased cars with higher accuracy than simple multiple regression or multivariate regression. SVM is better able to deal with very high dimensional data (number of features used to predict the price) and can avoid both over-fitting and underfitting. In particular, she used a genetic algorithm to find the optimal parameters for SVM in less time. The only drawback of this study is that the improvement of SVM regression over simple regression was not expressed in simple measures like mean deviation or variance. In another university thesis, Richardson working on the hypothesis that car manufacturers are more willing to produce vehicles which do not depreciate rapidly. In particular, by using a multiple regression analysis, he showed that hybrid cars (cars which use two different power sources to propel the car, i.e. they have both an internal combustion engine and an electric motor) are more able to keep their value than traditional vehicles. This is likely due to more environmental concerns about the climate and because of its higher fuel efficiency. The importance of other factors like age, mileage, make and MPG (miles per gallon) were also considered in this study. He collected all his data from various websites.

Wu et al. used neuro-fuzzy knowledge based system to predict the price of used cars. Only three factors namely: the make of the car, the year in which it was manufactured and the engine style were considered in this study. The proposed system produced similar results as compared to simple

regression methods. Car dealers in USA sell hundreds of thousands of cars every year through leasing. Most of these cars are returned at the end of the leasing period and must be resold. Selling these cars at the right price have major economic connotation for their success. In response to this, the ODAV (Optimal Distribution of Auction Vehicles) system was developed by Du et al. This system not only estimates a best price for reselling the cars but also provides advice on where to sell the car. Since the United States is a huge country, the location where the car is sold also has a non-trivial impact on the selling price of used cars. A k-nearest neighbour regression model was used for forecasting the price. Since this system was started in 2003, more than two million vehicles have been distributed via this system.

Gonggi proposed a new model based on artificial neural networks to forecast the residual value of private used cars. The main features used in this study were: mileage, manufacturer and estimate useful life. The model was optimized to handle nonlinear relationships which cannot be done with simple linear regression methods. It was found that this model was reasonably accurate in predicting the residual value of used cars.

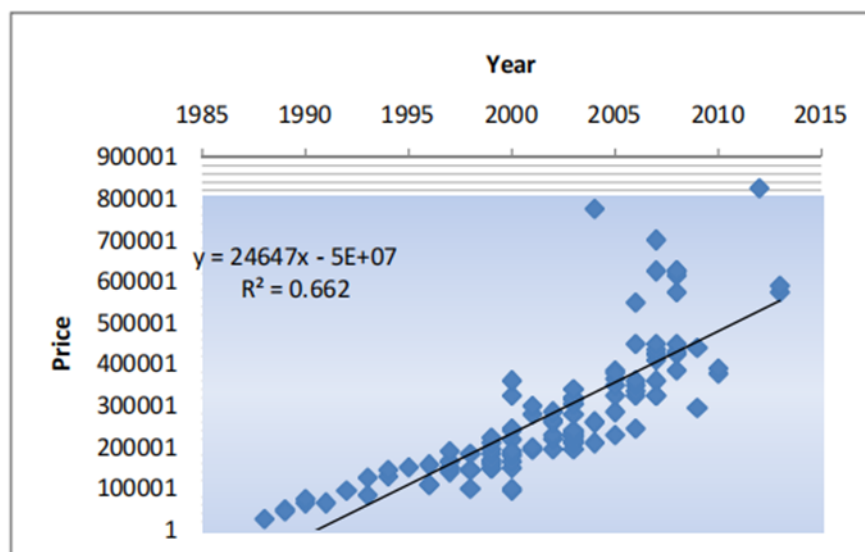
Methodology

Data was collected from <<petites annonces>> found in daily newspapers such as L'Express and Le Defi. We made sure that all the data was collected in less than one month interval as time itself could have an appreciable impact on the price of cars. In Mauritius, seasonal patterns is not really a problem as this does not really affect the purchase or selling of cars. The following data was collected for each car: make, model, volume of cylinder (funnily this is usually considered same as horsepower in Mauritius), mileage in km, year of manufacture, paint colour, manual/automatic and price. Only cars which had their price listed were recorded. Because many of the columns were sparse they were removed. Thus, paint colour and manual/automatic features were removed. The data was then further tweaked to remove records in which either the age (year) or the cylinder volume was not available. Model was also removed as it would have been extremely difficult to get enough records for all the variety of car models that exist. Although data for mileage was sparse, it was kept as it is considered to be a key factor in determining the price of used cars.

Implementation and Evaluation

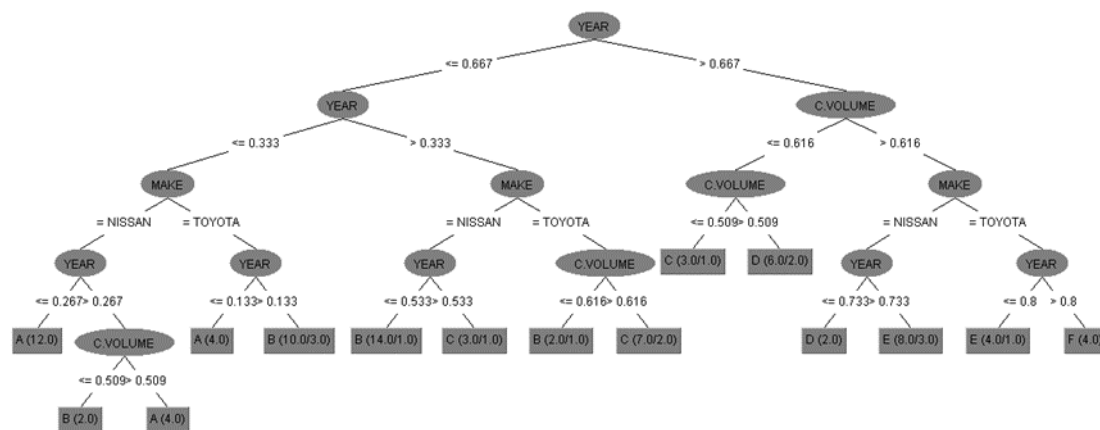
4.1. Multiple Linear Regression Analysis

The lack of mileage information for most of the cars did not allow us to use it to forecast the price. The Pearson correlation coefficient (r) was computed between different pairs of features. The value of r was found to be -0.33 between year and mileage, which means that there is a weak negative correlation between the year in which the car was manufactured and its mileage. The relation is not strong enough to use year information to predict mileage information. The relation between mileage and price is also very similar. On the other hand, there is a weak positive correlation between cylinder volume and price. This means that on average prices for cars with higher cylinder volume tends to be slightly higher. As anticipated, there is a very high correlation between the price of a car and the year in which it was manufactured. New cars have higher prices. Since all the three makes that we have considered are japanese manufacturers, we are assuming this will not impact on the price. There are many ups and down. In an attempt to increase the linearity, some outliers were removed. These were cars whose age were 20 and above or less than 4 years. The regression coefficient changed slightly from 0.814 to 0.819. Another variation was to use the logarithmic value of price instead of the actual values. In this case, the regression coefficient showed a significant increase from 0.814 to 0.851. In the first case, the mean deviation from the predicted price was Rs57468 while in the logarithmic scenario, it was at Rs51, 084. We can thus conclude that while logarithmic regression is slightly better than simple, linear regression on one variable alone is not sufficient to predict an accurate price for used cars. We also found that the regression coefficient is much higher for Nissan cars (0.917) than for toyota cars (0.803).



Decision Tree:

Only Nissan and Toyota cars were considered for building the decision tree. The prices were grouped into six nominal categories as most of the popular decision tree algorithms cannot handle numeric outputs. There are many gaps in the ranges that have been defined because there were no cars within these ranges although it is certainly possible to get new data which fits within these zones. These large gaps have been useful in determining the boundaries for the classes. The defined ranges can certainly be extended when more data is available. J48 is a free Java implementation of the C4.5 decision tree algorithm which is found in Weka. The following attributes were used to build the tree: make, model, cylinder volume, year and price.



J48 Decision Tree on Training Set Without Car Model

Naïve Bayes

Naïve Bayes (NaiveBayes in Weka) is one of the most useful machine learning technique [14]. There are two reasons for that. Firstly, it is very easy to implement in software and secondary the accuracy is usually as good as more complex algorithms. Two different experiments were conducted. In the first one, the original data was used, i.e., the attributes' values were not converted in any way. The same attributes were used as in decision trees.

K-Nearest Neighbours (kNN)

K-nearest neighbour (IBk in Weka) is a machine learning technique in which the

new (unknown) data is compared to all the existing records in order to locate the best match(es). Despite its apparent simplicity, a lot of care has to be taken in pre processing the data otherwise we can easily go off-track. Only three attributes were considered namely the make, year and cylinder volume. However, the data set was split into different sets, one containing only Toyota cars and the other only Nissan cars. This was done because most software cannot handle nominal values appropriately but this allowed us to compare the performance on each make. In general, it was also found that Toyota cars of the same age and cylinder are more expensive than Nissan cars with the same features. The data for year and cylinder had to be normalized to prevent large values (from one feature) from over-shadowing smaller values (from another feature). Thus, the following formulae were applied for normalizing the data. There were two options for normalizing the year.

1. Normalized value for Year = $(\text{Year of Manufacture} - 2014)/19 + 1$

2. Normalized value for Year = $(\text{Year of Manufacture} - 2010)/15 + 1$

The second option was chosen because most used cars are more than 4 years old. Used cars which are less than 4 years old are very likely to be outliers and may significantly affect the prediction performance. The divisor is the span of years from the newest car to the oldest car in the database. One (1) is then added to the quotient to bring the value between 0 and 1 and to make sure that newer cars have higher normalized values than older cars.

The formula for normalizing cylinder volume is as follows:

Normalized value = $\text{Cylinder Volume} / \text{Maximum}(\text{Cylinder Volume})$

A simple formula was used to normalize the data for cylinder volume. We have to ensure that cars with higher values for cylinder volume have higher normalized values than cars with lower values.

This seems to suggest that prices for Nissan cars are more consistent than prices for Toyota cars. Also, we can see that for Toyota cars, the best value of k is 1 and the performance degrades for increasing values of k while for Nissan cars, the best value of k is 5. But performance is almost the same even for lower values of k . Performance starts to degrade for higher values of k .

Evaluation and Conclusion

In this paper, four different machine learning techniques have been used to forecast the price of used cars in Mauritius. The mean error with linear regression was about Rs51, 000 while for kNN it was about Rs27, 000 for Nissan cars and about Rs45, 000 for Toyota cars. J48 and NaiveBayes accuracy dangled between 60-70% for different combinations of parameters. The main weakness of decision trees and naïve bayes is their inability to handle output classes with numeric values. Hence, the price attribute had to be classified into classes which contained a range of prices but this evidently introduced further grounds for inaccuracies. The main limitation of this study is the low number of records that have been used. As future work, we intend to collect more data and to use more advanced techniques like artificial neural networks, fuzzy logic and genetic algorithms to predict car prices.

2.3. Predicting the Price of Second-hand Cars using Artificial Neural Networks

Abstract

The number of cars on Mauritian roads has been rising consistently by 5% during the last decade. In 2014, 173 954 cars were registered at the National Transport Authority. Thus, one Mauritian in every six owns a car, most of which are second hand reconditioned cars and used cars. The aim of this study is to assess whether it is possible to predict the price of second-hand cars using artificial neural networks. Thus, data for 200 cars from different sources was gathered and fed to four different machine learning algorithms. We found that support vector machine regression produced slightly better results than using a neural network or linear regression. However, some of the predicted values are quite far away from the actual prices, especially for higher priced cars. Thus, more investigations with a larger data set are required and more experimentation with different network type and structures is still required in order to obtain better predictions.

INTRODUCTION

According to the data obtained from the National Transport Authority (2014), there has been an increase of 254% in the number of cars from 2003 (68, 524) to 2014 (173, 954). We can thus infer that the sale of second-hand imported (reconditioned) cars and second-hand used cars has eventually increase given that new cars represent only a very small percentage of the total number of cars sold each year. Most individuals in Mauritius who buy new cars also want to know about the resale value of their cars after some years so that they can sell it in the used car market. Price

prediction of second-hand cars depends on numerous factors. The most important ones are manufacturing year, make, model, mileage, horsepower and country of origin. Some other factors are type and amount of fuel per usage, the type of braking system, its acceleration, the interior style, its physical state, volume of cylinders (measured in cubic centimeters), size of the car, number of doors, weight of the car, consumer reviews, paint colour and type, transmission type, whether it is a sports car, sound system, cosmic wheels, power steering, air conditioner, GPS navigator, safety index etc. In the Mauritian context, there are some special factors that are also usually considered such as who were the previous owners and whether the car has had any serious accidents. Thus, predicting the price of second-hand cars is a very laudable enterprise. In this paper, we will assess whether neural networks can be used to accurately predict the price of secondhand cars. The results will also be compared with other methods like linear regression and support vector regression. This paper proceeds as follows. In Section II, various works on neural networks and price prediction have been summarised. The methodology and data collection are described in section III. Section IV presents the results for price prediction of second-hand cars. Finally, we end the paper with a conclusion and some ideas towards future works.

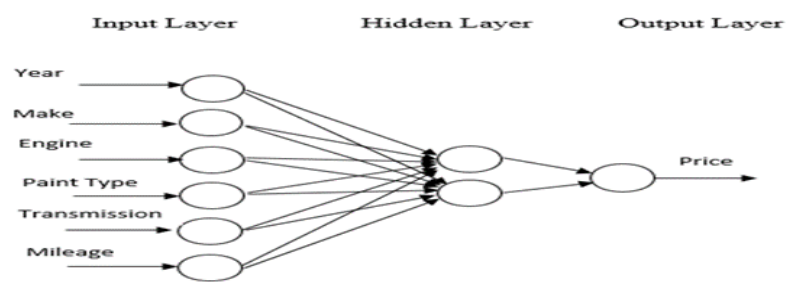
RELATED WORKS

Predicting the price of second-hand cards has not received much attention from academia despite its huge importance for the society. Bharambe and Dharmadhikari (2015) used artificial neural networks (ANN) to analyse the stock market and predict market behaviour. They claimed that their proposed approach is more accurate than existing ones by 25%. Pudaruth (2014) used four different supervised machine learning techniques namely kNN (k-Nearest Neighbour), Naïve Bayes, linear regression and decision trees to predict the price of second-hand cars. The best result was obtained using kNN which had a mean error of 27000 rupees. Jassbi et al. (2011) used two different neural networks and regression methods to predict the thickness of paint coatings on cars. The error for the final thickness of the paint was found to be 2/99 microns for neural networks and 17/86 for regression. Ahangar et al. (2010) also compared the use of neural networks with linear regression in order to predict the stock prices of companies in Iran. They also found that neural networks had superior performance both in terms of accuracy and speed compared to linear regression. Listiani (2009) used support vector machines (SVM) to predict the price of leased cars. They showed that SVM performed better than simple linear regression and multivariate regression. Iseri and Karlik (2009) used neural networks to predict the price of automobiles and achieved a mean square error of 8% compared with 14.4% for regression. Yeo (2009) used neural networks to predict the retention

rate for policy holders of automobile insurance. The neural network was able to predict which customers were likely to renew their policy and which ones would terminate soon. Doganis et al. (2006) used artificial neural networks and genetic algorithm in order to predict the sales of fresh milk with an accuracy of 95.4%. Rose (2003) used neural networks to predict the production of cars for different manufacturers. Thus, we have seen that neural networks have been used successfully for predicting the price of various commodities. Our objective, therefore, in this work, is to use neural networks in a new application, i.e., that of predicting the price of second-hand cars.

METHODOLOGY

In order to carry out this study, data have been obtained from different car websites and from the small adverts sections found in daily newspapers like L'Express and Le Defi. The data was collected in less than one month interval (i.e. in the month of August in 2014) because like other goods, the price of cars also changes with time. Two hundred records were collected. The data comprises of different features for second-hand cars such as the year (YEAR) in which it was manufactured, the make (MAKE), engine capacity (ENGINE) measured in cubic centimetres, paint (PAINT) type (normal or metallic), transmission (T/N) type (manual or automatic), mileage (MILEAGE) (number of kilometres the car has been driven) and its price (PRICE) in Mauritian rupees. eleven records selected from our dataset of 200 records. The range for the year attribute was 2000-2012. A total of fifteen make was studied. Chevrolet and Peugeot had only 3 instances while Toyota has 63 instances. The smallest horsepower in the dataset was 900 and the highest one was 2900. For paint type, 0 stood for normal paint while 1 stood for metallic. For transmission type, a value of 0 means manual transmission while a value of 1 means automatic transmission. The lowest mileage recorded was 2000 km and the highest one was 275,000 km while the price range was 110000 to 685000.



Neural Network Architecture

A neural network with six inputs and one hidden layer containing two nodes was used to predict the price of second-hand cars.

EXPERIMENTS AND RESULTS

A large number of experiments have been conducted in order to find the best network structure and the best parameters for the neural network. We found that a neural network with 1 hidden layer and 2 nodes produced the smallest mean absolute error among various neural network structures that were experimented with. However, we found that Support Vector Regression and a multilayer perceptron with back-propagation produced slightly better predictions than linear regression while the k-Nearest Neighbour algorithm had the worst accuracy among these four approaches. All experiments were performed with a cross validation value of 10 folds. Pudaruth (2014) used only 97 records, 3 make and only 3 features and obtained a mean absolute error of 27000 with kNN (k=1). However, in this work, we have used 200 records with 6 inputs and experimented with more complex approaches. Although, the mean absolute error was slightly higher in our experiments, the value of 30605 can be considered to be a satisfactory outcome as the mean price of the cars was found to be Rs 311586, which is less than 10%. The actual price values obtained from the different sources have been assumed to reflect the true value of the cars but we should point out that these values are estimated by car owners who often do not have much experience in the car business. the variation in actual price (red line) against the variation in the predicted predicted price (green line) using a multi-layer perceptron. The graph shows that as the price gets higher, the deviation from the actual price also increases by a small amount. Nevertheless, the graph also shows that the predictions are also fairly accurate and can be relied upon in many cases.

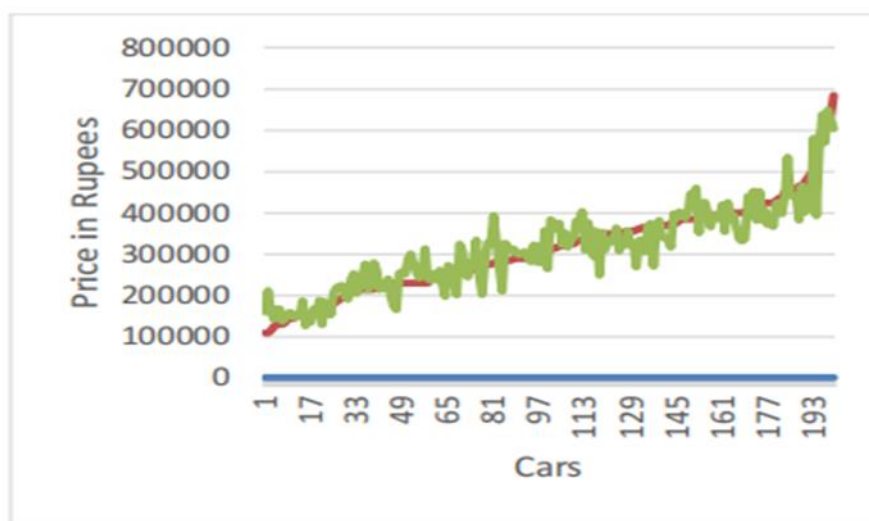


Figure 3. Actual Price v/s Predicted Price using MLP

CONCLUSION:

The aim of this paper was to predict the price of second-hand reconditioned and secondhand used cars in Mauritius. The car market has been increasing steadily by around 5% for the last ten years, showing the high demand for cars by the Mauritian population. There are hundreds of car websites in Mauritius but none of them provide such a facility to predict the price of used cars based on their attributes. Our dataset of 200 records was used with the cross-validation technique with ten folds. The car make, year manufactured, paint type, transmission type, engine capacity and mileage have been used to predict the price of second-hand cars using four different machine learning algorithms. The average residual value was reasonably low for all four approaches. Thus, we conclude that predicting the price of second-hand cars is a very risky enterprise but which is feasible. This system will be very useful to car dealers and car owners who need to assess the value of their cars. In the future, we intend to collect more data and more features and to use a larger variety of machine learning algorithms to do the prediction.

2.4 USED CAR PRICE PREDICTION

Abstract :

The price of a new car in the industry is fixed by the manufacturer with some additional costs incurred by the Government in the form of taxes. So, customers buying a new car can be assured of the money they invest to be worthy. But, due to the increased prices of new cars and the financial incapability of the customers to buy them, Used Car sales are on a global increase. Therefore, there is an urgent need for a Used Car Price Prediction system which effectively determines the worthiness of the car using a variety of features. Existing System includes a process where a seller decides a price randomly and buyer has no idea about the car and it's value in the present day scenario. In fact, seller also has no idea about the car's existing value or the price he should be selling the car at. To overcome this problem we have developed a model which will be highly effective. Regression Algorithms are used because they provide us with continuous value as an output and not a

categorized value. Because of which it will be possible to predict the actual price a car rather than the price range of a car. User Interface has also been developed which acquires input from any user and displays the Price of a car according to user's inputs.

INTRODUCTION

Determining whether the listed price of a used car is a challenging task, due to the many factors that drive a used vehicle's price on the market. The focus of this project is developing machine learning models that can accurately predict the price of a used car based on its features, in order to make informed purchases. We implement and evaluate various learning methods on a dataset consisting of the sale prices of different makes and models. We will compare the performance of various machine learning algorithms like Linear Regression, Ridge Regression, Lasso Regression, Elastic Net, Decision Tree Regressor and choose the best out of it. Depending on various parameters we will determine the price of the car. Regression Algorithms are used because they provide us with continuous value as an output and not a categorized value because of which it will be possible to predict the actual price a car rather than the price range of a car. User Interface has also been developed which acquires input from any user and displays the Price of a car according to user's inputs.

Literature Survey

The first paper is Predicting the price of Used Car Using Machine Learning Techniques. In this paper, they investigate the application of supervised machine learning techniques to predict the price of used cars in Mauritius. The predictions are based on historical data collected from daily newspapers. Different techniques like multiple linear regression analysis, k-nearest neighbours, naïve bayes and decision trees have been used to make the predictions. The Second paper is Car Price Prediction Using Machine Learning Techniques. Considerable number of distinct attributes are examined for the reliable and accurate prediction. To build a model for predicting the price of used cars in Bosnia and Herzegovina, they have applied three machine learning techniques (Artificial Neural Network, Support Vector Machine and Random Forest). The Third paper is Price Evaluation model in second hand car system based on BP neural networks. In this paper, the price evaluation model based on

big data analysis is proposed, which takes advantage of widely circulated vehicle data and a large number of vehicle transaction data to analyze the price data for each type of vehicles by using the optimized BP neural network algorithm. It aims to establish a second-hand car price evaluation model to get the price that best matches the car.

1.2 Requirements

- Hardware requirements
- Operating system- Windows 7,8,10
- Processor- dual core 2.4 GHz (i5 or i7 series Intel processor or equivalent AMD)
- RAM-4GB
- Software Requirements
- Python
- Pycharm
- PIP 2.7
- Jupyter Notebook
- Chrome

METHODOLOGY

There are two primary phases in the system:

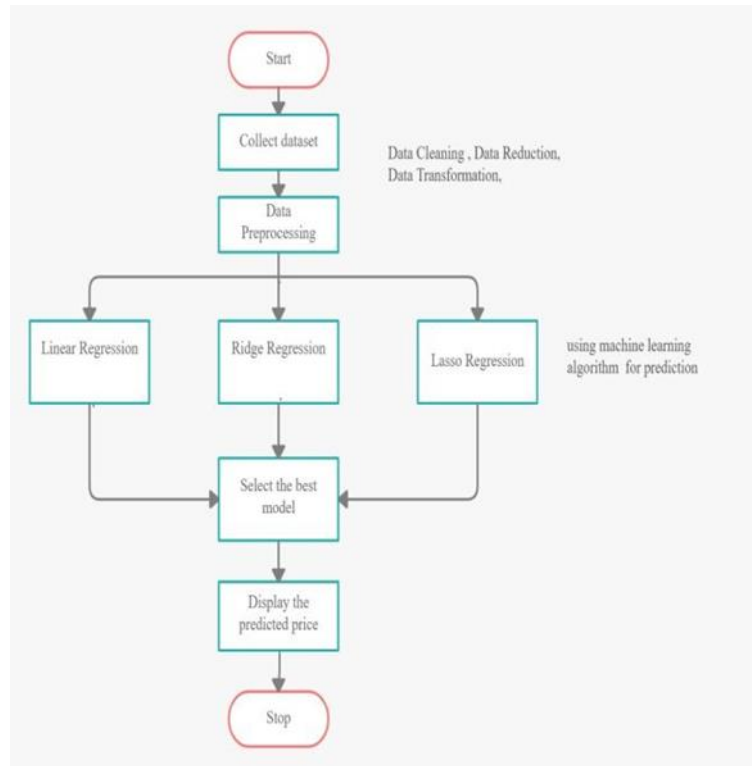
1. Training phase: The system is trained by using the data in the data set and fits a model (line/curve) based on the algorithm chosen accordingly.
2. Testing phase: the system is provided with the inputs and is tested for its working. The accuracy is checked. And therefore, the data that is used to train the model or test it, has to be appropriate. The system is designed to detect and predict price of used car and hence appropriate algorithms must be used to do the two different tasks. Before the algorithms are selected for further use, different algorithms were compared for its accuracy. The well-suited one for the task was chosen.

Objective

- To develop a efficient and effective model which predicts the price of a used car according to user's inputs.

- To achieve good accuracy.
- To develop a User Interface(UI) which is user-friendly and takes input from the user and predicts the price.

PROPOSED SYSTEM



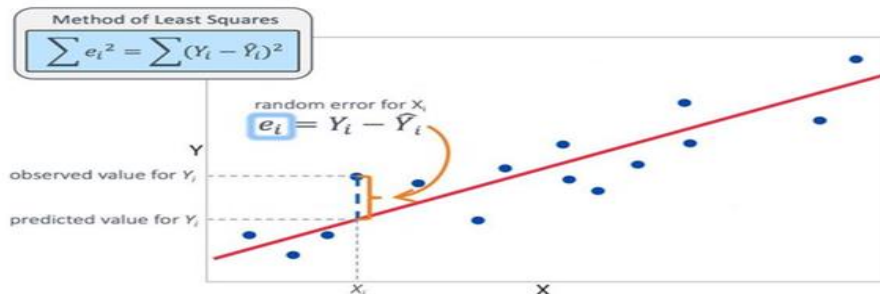
Proposed System Flowchart

As shown in the above figure, the process starts by collecting the dataset. The next step is to do Data Preprocessing which includes Data cleaning, Data reduction, Data Transformation. Then, using various machine learning algorithms we will predict the price. The algorithms involve Linear Regression, Ridge Regression and Lasso Regression. The best model which predicts the most accurate price is selected. After selection of the best model the predicted price is displayed to the user according to user's inputs. User can give input through website to for used car price prediction to machine learning model.

Linear Regression

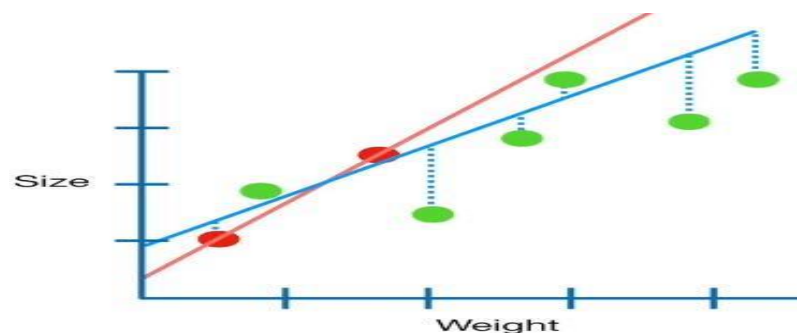
Linear Regression attempt to model the relationship between two variables by fitting a linear equation to observed data. The other is considered to be dependent variable. For Example:

A modeler might want to relate weights of individuals to their heights using a linear regression model.



Ridge Regression

A Ridge regressor is basically a regularized version of Linear Regressor. The regularized term has the parameter 'alpha' which controls the regularization of the model i.e helps in reducing the variance of the estimates.



Lasso Regression

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

legitimate URL. By analyzing the "on Mouse Over" events in the web page we can determine if such a modification has occurred.

FUTURE SCOPE

In future this machine learning model may bind with various website which can provide real time data for price prediction. Also we may add large historical data of car price which can help to improve accuracy of the machine learning model. We can build an android app as user interface for interacting with user. For better performance, we plan to judiciously design deep learning network structures, use adaptive learning rates and train on clusters of data rather than the whole dataset. consists of three main files: manifest.json, content.js, background.js. The manifest file provides all the meta data information about the extension to Chrome browser. Addition-ally, it also specifies all the files and other resources associated to the extension. The content.js file loads on every page in the Chrome browser, post the extension deployment. However, it is an unprivileged module, which has direct access only to the DOM elements and needs supporting files to interact to external APIs and browser user interface manipulation. The supplementary file background.js aids the content script with these interactions, which is termed as message passing.

CONCLUSION

The increased prices of new cars and the financial incapability of the customers to buy them, Used Car sales are on a global increase. Therefore, there is an urgent need for a Used Car Price Prediction system which effectively determines the worthiness of the car using a variety of features. The proposed system will help to determine the accurate price of used car price prediction. This paper compares 3 different algorithms for machine learning : Linear Regression, Lasso Regression and Ridge Regression.

Chapter 3

3.1 EXISTING SYSTEM:

- Forecasting Based on Your Current Funnel
- Forecasting Based on Lead Scores and Multiple Variables
- Lead Scoring Tools
- Qualitative techniques
- Time series analysis and projection
- Casual Models

PROBLEM STATEMENT:

- We need to have historical data for a long time period to capture seasonality. However, often we do not have historical data for a target variable, for example in case when a new product is launched. At the same time we have sales time series for a similar product and we can expect that our new product will have a similar sales pattern.
- Sales data can have a lot of outliers and missing data.

- We must clean outliers and interpolate data before using a time series approach.
- We need to take into account a lot of exogenous factors like covid 19 which have impact on sales.

3.2 PROPOSED SYSTEM:

- For our analysis, we used sales historical data from Kaggle . The calculations were conducted in the Python environment using the main packages pandas, sklearn, numpy, keras, matplotlib, seaborn. To conduct the analysis, Pycharm IDE was used.
- we conducted the descriptive analytics, which is a study of sales distributions, data visualization with different pairplots. It is helpful in finding correlations and sales drivers on which we focus.
- A specific feature of most machine-learning methods is that they can work with stationary data only. Let us consider the supervised machine-learning approach using sales historical time series. As covariates, we used categorical features: promo, day of week, day of month, month. For categorical features, we applied one-hot encoding, when one categorical variable was replaced by n binary variables, where n is the amount of unique values of categorical variables.
- Linear Regression Model is used to forecast the sales

Advantage:

- It has been growing tremendously and have a very bright future.
- it can easily predict using before data.
- the condition is, in real life, there are many factors affecting the production and sale of a company.
- It is very useful for future sales for company.
- The use of regression approaches for sales forecasting can often give us better results compared to time series methods.
- The accuracy on the validation set is an important indicator for choosing an optimal number of iterations of machine-learning algorithms.
- Can be used to make sales prediction when there is a small number of historical data for specific sales time series.

Fig 3.1 Block Diagram

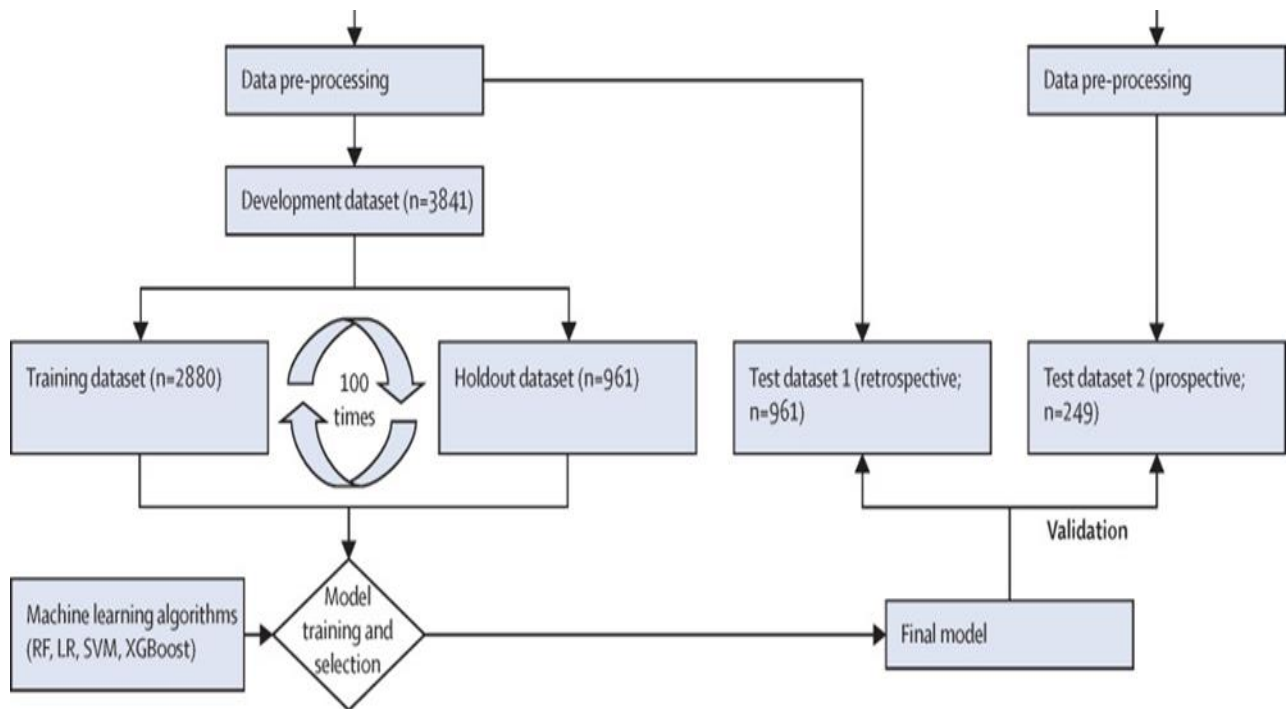
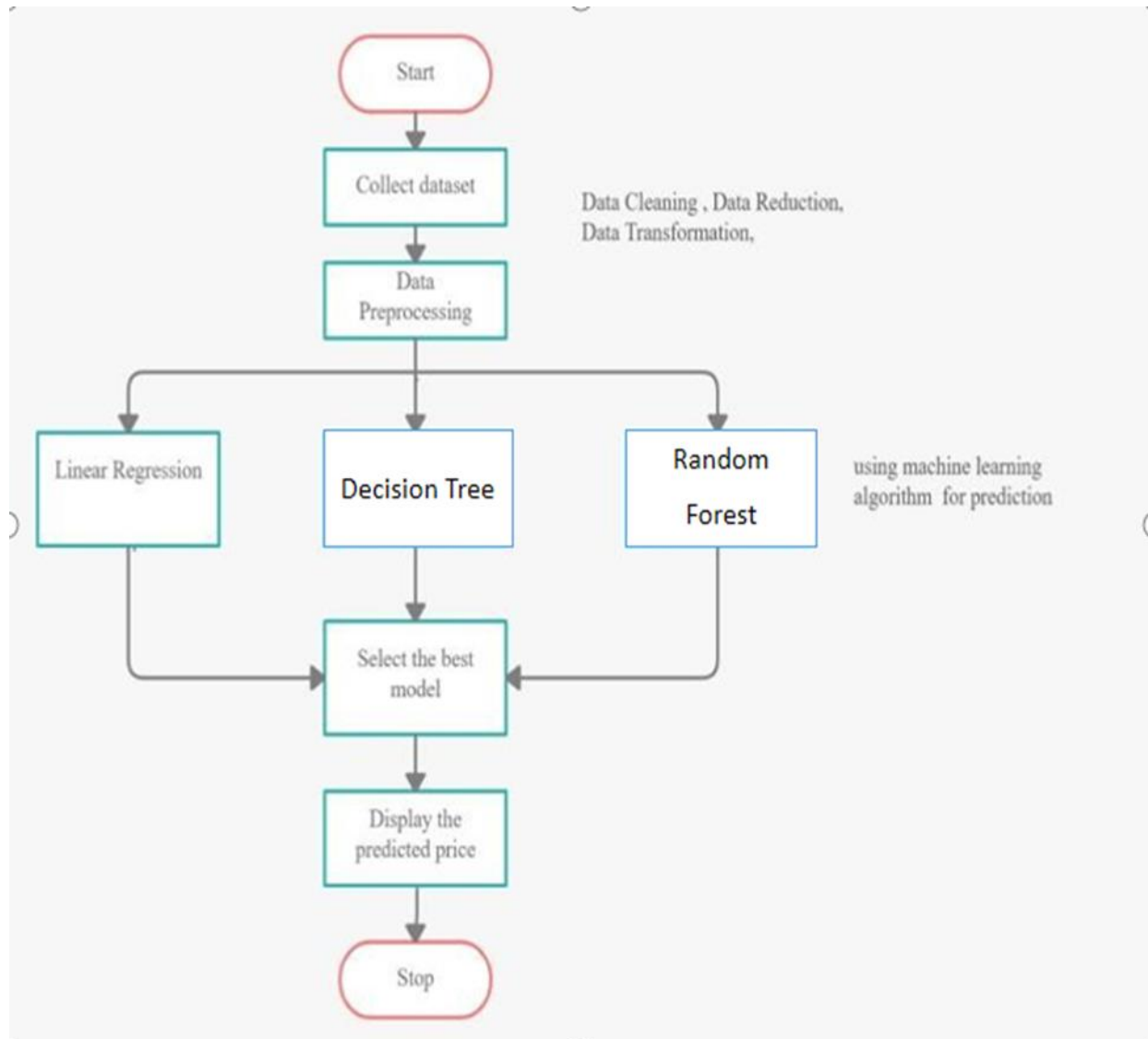


Fig 3.2 Flow Diagram



3.4 DECISION TREE ALGORITHM:

Introduction:

Till now we have learned about linear regression, logistic regression, and they were pretty hard to understand. Let's now start with Decision tree's and I assure you this is probably the easiest algorithm in Machine Learning. There's not much mathematics involved here. Since it is very

easy to use and interpret it is one of the most widely used and practical methods used in Machine Learning.

What is a Decision Tree?

It is a tool that has applications spanning several different areas. Decision trees can be used for classification as well as regression problems. The name itself suggests that it uses a flowchart like a tree structure to show the predictions that result from a series of feature-based splits. It starts with a root node and ends with a decision made by leaves. Before learning more about decision trees let's get familiar with some of the terminologies.

Root Nodes – It is the node present at the beginning of a decision tree from this node the population starts dividing according to various features.

Decision Nodes – the nodes we get after splitting the root nodes are called Decision Node

Leaf Nodes – the nodes where further splitting is not possible are called leaf nodes or terminal nodes

Sub-tree – just like a small portion of a graph is called sub-graph similarly a sub-section of this decision tree is called sub-tree.

Pruning – is nothing but cutting down some nodes to stop overfitting.

Example of a decision tree.

Let's understand decision trees with the help of an example.

Decision trees are upside down which means the root is at the top and then this root is split into various several nodes. Decision trees are nothing but a bunch of if-else statements in layman terms. It checks if the condition is true and if it is then it goes to the next node attached to that decision.

In the below diagram the tree will first ask what is the weather? Is it sunny, cloudy, or rainy? If yes then it will go to the next feature which is humidity and wind. It will again check if there is a strong wind or weak, if it's a weak wind and it's rainy then the person may go and play. Did you notice anything in the above flowchart? We see that if the weather is cloudy then we must go to play. Why didn't it split more? Why did it stop there? To answer this question, we need to know about few more concepts like entropy, information gain, and Gini index. But in simple terms, I can say here that the output for the training dataset is always yes for cloudy weather, since there is no disorderliness here we don't need to split the node further. The goal of machine learning is to decrease uncertainty or disorders from the dataset and for this, we use decision trees. Now you must be thinking how do I know what should be the root node? what should be the decision node? when should I stop splitting? To decide this, there is a metric called "Entropy" which is the amount of uncertainty in the dataset.

Entropy

Entropy is nothing but the uncertainty in our dataset or measure of disorder. Let me try to explain this with the help of an example.

Suppose you have a group of friends who decides which movie they can watch together on Sunday. There are 2 choices for movies, one is "**Lucy**" and the second is "**Titanic**" and now everyone has to tell their choice. After everyone gives their answer we see that "Lucy" gets 4 votes and "Titanic" gets 5 votes. Which movie do we watch now? Isn't it hard to choose 1 movie now because the votes for both the movies are somewhat equal. This is exactly what we call disorderliness, there is an equal number of votes for both the movies, and we can't really decide which movie we should watch. It would have been much easier if the votes for "Lucy" were 8 and for "Titanic" it was 2. Here we could easily say that the majority of votes are for "Lucy" hence everyone will be watching this movie.

In a decision tree, the output is mostly “yes” or “no”

The formula for Entropy is shown below:

Here p_+ is the probability of positive class p_- is the probability of negative class

S is the subset of the training example

How do Decision Trees use Entropy?

Now we know what entropy is and what is its formula, Next, we need to know that how exactly does it work in this algorithm. Entropy basically measures the impurity of a node. Impurity is the degree of randomness; it tells how random our data is. A **pure sub-split** means that either you should be getting “yes”, or you should be getting “no”.

Suppose feature 1 had 8 yes and 4 no, after the split feature 2 get 5 yes and 2 no whereas feature 3 gets 3 yes and 2 no. We see here the split is not pure, why? Because we can still see some negative classes in both the feature. In order to make a decision tree, we need to calculate the impurity of each split, and when the purity is 100% we make it as a leaf node. To check the impurity of feature 2 and feature 3 we will take the help for Entropy formula. For feature 2 the entropy is as follows:

For feature 3, We can clearly see from the tree itself that feature 2 has low entropy or more purity than feature 3 since feature 2 has more “yes” and it is easy to make a decision here.

Always remember that the higher the Entropy, the lower will be the purity and the higher will be the impurity. As mentioned earlier the goal of machine learning is to decrease the uncertainty or impurity in the dataset, here by using the entropy we are getting the impurity of a feature or a particular node, we don't know if the parent entropy or the entropy of a particular node has decreased or not.

For this, we bring a new metric called “Information gain” which tells us how much the parent entropy has decreased after splitting it with some feature.

Information Gain

Information gain measures the reduction of uncertainty given some feature and it is also a deciding factor for which attribute should be selected as a decision node or root node. It is just entropy of the full dataset – entropy of the dataset given some feature.

When to stop splitting?

You must be asking this question to yourself that when do we stop growing our tree? Usually, real-world datasets have a large number of features, which will result in a large number of splits, which in turn gives a huge tree. Such trees take time to build and can lead to overfitting. That means the tree will give very good accuracy on the training dataset but will give bad accuracy in test data.

There are many ways to tackle this problem through hyperparameter tuning. We can set the maximum depth of our decision tree using the **max_depth** parameter. The more the value of **max_depth**, the more complex your tree will be. The training error will of-course decrease if we increase the **max_depth** value but when our test data comes into the picture, we will get a very bad accuracy. Hence you need a value that will not overfit as well as underfit our data and for this, you can use GridSearchCV.

Another way is to set the minimum number of samples for each split. It is denoted by

min_samples_split.

Here we specify the minimum number of samples required to do a split. For example, we can use a minimum of 10 samples to reach a decision. That means if a node has less than 10 samples then using this parameter, we can stop the further splitting of this node and make it a leaf node.

There are more hyperparameters such as :

min_samples_leaf – represents the minimum number of samples required to be in the leaf node. The more you increase the number, the more is the possibility of overfitting.

max_features – it helps us decide what number of features to consider when looking for the best

Pruning:

It is another method that can help us avoid overfitting. It helps in improving the performance of the tree by cutting the nodes or sub-nodes which are not significant. It removes the branches which have very low importance.

There are mainly 2 ways for pruning:

Pre-pruning – we can stop growing the tree earlier, which means we can prune/remove/cut a node if it has low importance **while growing** the tree.

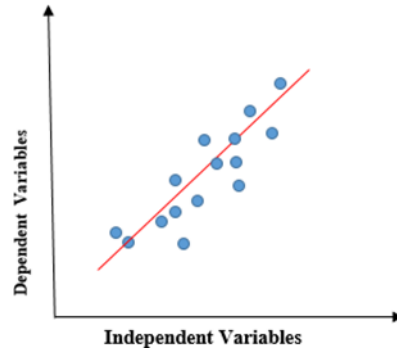
Post-pruning – once our **tree is built to its depth**, we can start pruning the nodes based on their significance.

Endnotes

To summarize, in this article we learned about decision trees. On what basis the tree splits the nodes and how to can stop overfitting. why linear regression doesn't work in the case of classification problems. In the next article, I will explain Random forests, which is again a new technique to avoid overfitting.

Linear Regression:

Linear regression is a quiet and simple statistical regression method used for predictive analysis and shows the relationship between the continuous variables. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression. If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called multiple linear regression. The linear regression model gives a sloped straight line describing the relationship within the variables.



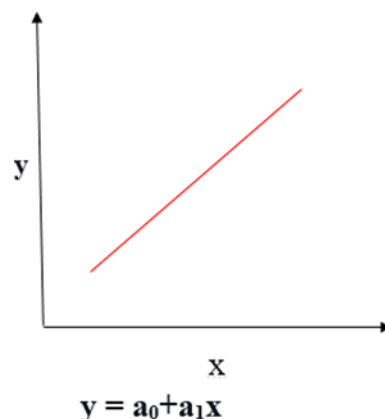
The above graph presents the linear relationship between the dependent variable and independent variables. When the value of x (independent variable) increases, the value of y (dependent variable) is likewise increasing. The red line is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best.

Let's say we want to estimate the salary of an employee based on year of experience. You have the recent company data, which indicates that the relationship between experience and salary. Here year of experience is an independent variable, and the salary of an employee is a dependent variable, as the salary of an employee is dependent on the experience of an employee. Using this insight, we can predict the future salary of the employee based on current & past information.

A regression line can be a Positive Linear Relationship or a Negative Linear Relationship.

Positive Linear Relationship

If the dependent variable expands on the Y-axis and the independent variable progress on X-axis, then such a relationship is termed a Positive linear relationship.



Negative Linear Relationship:

If the dependent variable decreases on the Y-axis and the independent variable increases on the X-axis, such a relationship is called a negative linear relationship. The goal of the linear regression algorithm is to get the best values for a_0 and a_1 to find the best fit line. The best fit line should have the least error means the error between predicted values and actual values should be minimized.

Cost function:

The cost function helps to figure out the best possible values for a_0 and a_1 , which provides the best fit line for the data points.

Cost function optimizes the regression coefficients or weights and measures how a linear regression model is performing. The cost function is used to find the accuracy of the mapping function that maps the input variable to the output variable. This mapping function is also known as the Hypothesis function.

In Linear Regression, Mean Squared Error (MSE) cost function is used, which is the average of squared error that occurred between the predicted values and actual values.

By simple linear equation $y=mx+b$ we can calculate MSE as:

Let's y = actual values, y_i = predicted values

Using the MSE function, we will change the values of a_0 and a_1 such that the MSE value settles at the minima. Model parameters x_i , b (a_0 , a_1) can be manipulated to minimize the cost function. These parameters can be determined using the gradient descent method so that the cost function value is minimum.

Gradient descent

Gradient descent is a method of updating a_0 and a_1 to minimize the cost function (MSE). A regression model uses gradient descent to update the coefficients of the line (a_0 , $a_1 \Rightarrow x_i$, b)

by reducing the cost function by a random selection of coefficient values and then iteratively update the values to reach the minimum cost function.

Imagine a pit in the shape of U. You are standing at the topmost point in the pit, and your objective is to reach the bottom of the pit. There is a treasure, and you can only take a discrete number of steps to reach the bottom. If you decide to take one footstep at a time, you would eventually get to the bottom of the pit but, this would take a longer time. If you choose to take longer steps each time, you may get to sooner but, there is a chance that you could overshoot the bottom of the pit and not near the bottom. In the gradient descent algorithm, the number of steps you take is the learning rate, and this decides how fast the algorithm converges to the minima. To update a_0 and a_1 , we take gradients from the cost function. To find these gradients, we take partial derivatives for a_0 and a_1 .

The partial derivatives are the gradients, and they are used to update the values of a_0 and a_1 . Alpha is the learning rate.

Impact of different values for learning rate

The blue line represents the optimal value of the learning rate, and the cost function value is minimized in a few iterations. The green line represents if the learning rate is lower than the optimal value, then the number of iterations required high to minimize the cost function. If the learning rate selected is very high, the cost function could continue to increase with iterations and saturate at a value higher than the minimum value, that represented by a red and black line.

Use case

In this, I will take random numbers for the dependent variable (salary) and an independent variable (experience) and will predict the impact of a year of experience on salary.

Random Forest

Introduction

Random forest is a **Supervised Machine Learning Algorithm** that is **used widely in Classification and Regression problems**. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. It performs better results for classification problems.

Real Life Analogy

Let's dive into a real-life analogy to understand this concept further. A student named X wants to choose a course after his 10+2, and he is confused about the choice of course based on his skill set. So he decides to consult various people like his cousins, teachers, parents, degree students, and working people. He asks them varied questions like why he should choose, job opportunities with that course, course fee, etc. Finally, after consulting various people about the course he decides to take the course suggested by most of the people.

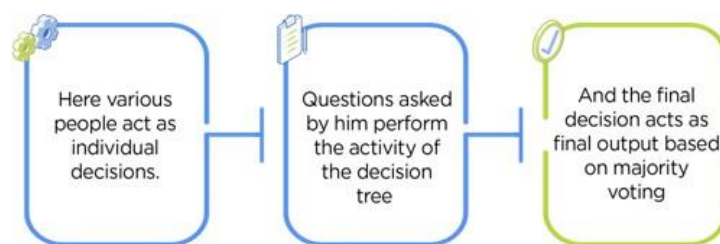


Fig 3.3 Real Life Analogy

Working of Random Forest Algorithm

Before understanding the working of the random forest we must look into the ensemble technique. Ensemble simply means combining multiple models. Thus a collection of models is used to make predictions rather than an individual model.

Ensemble uses two types of methods:

Bagging– It creates a different training subset from sample training data with replacement & the final output is based on majority voting. For example, Random Forest.

Boosting– It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. For example, ADA BOOST, XG BOOST.

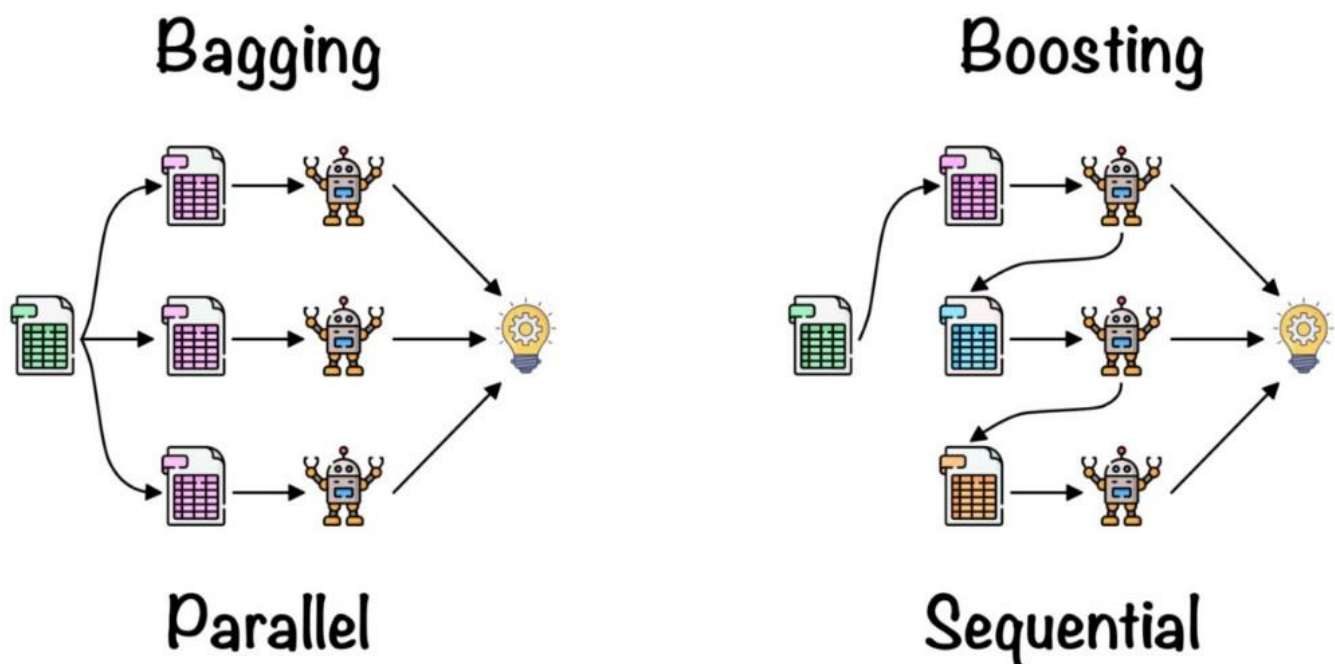
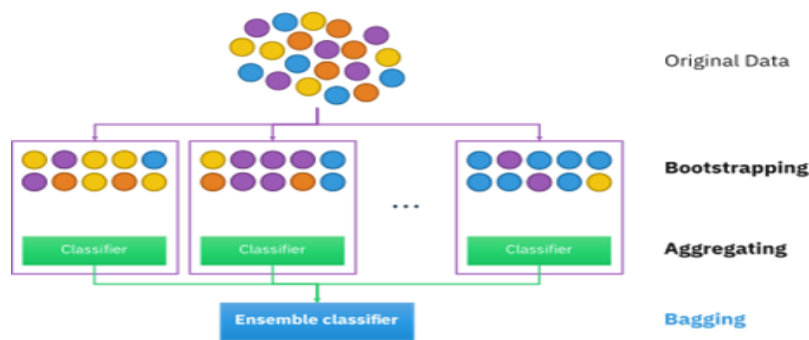


Fig 3.4 Bagging Parallel & Boosting Sequential

As mentioned earlier, Random forest works on the Bagging principle. Now let's dive in and understand bagging in detail.

Bagging

Bagging, also known as Bootstrap Aggregation is the ensemble technique used by random forest. Bagging chooses a random sample from the data set. Hence each model is generated from the samples (Bootstrap Samples) provided by the Original Data with replacement known as row sampling. This step of row sampling with replacement is called bootstrap. Now each model is trained independently which generates results. The final output is based on majority voting after combining the results of all models. This step which involves combining all the results and generating output based on majority voting is known as aggregation.



Now let's look at an example by breaking it down with the help of the following figure. Here the bootstrap sample is taken from actual data (Bootstrap sample 01, Bootstrap sample 02, and Bootstrap sample 03) with a replacement which means there is a high possibility that each sample won't contain unique data.

Now the model (Model 01, Model 02, and Model 03) obtained from this bootstrap sample is trained independently. Each model generates results as shown. Now Happy emoji is having a majority when compared to sad emoji. Thus based on majority voting final output is obtained as Happy emoji.

Steps involved in random forest algorithm:

Step 1: In Random forest n number of random records are taken from the data set having k number of records.

Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output.

Step 4: Final output is considered based on Majority Voting or Averaging for Classification and regression respectively.

For example: consider the fruit basket as the data as shown in the figure below. Now n number of samples are taken from the fruit basket and an individual decision tree is constructed for each sample. Each decision tree will generate an output as shown in the figure. The final output is

considered based on majority voting. In the below figure you can see that the majority decision tree gives output as an apple when compared to a banana, so the final output is taken as an apple.

Important Features of Random Forest

1. **Diversity**- Not all attributes/variables/features are considered while making an individual tree, each tree is different.
2. **Immune to the curse of dimensionality**- Since each tree does not consider all the features, the feature space is reduced.
3. **Parallelization**-Each tree is created independently out of different data and attributes. This means that we can make full use of the CPU to build random forests.
4. **Train-Test split**- In a random forest we don't have to segregate the data for train and test as there will always be 30% of the data which is not seen by the decision tree.
5. **Stability**- Stability arises because the result is based on majority voting/ averaging.

Difference Between Decision Tree & Random Forest

Random forest is a collection of decision trees; still, there are a lot of differences in their behavior.

Decision trees	Random Forest
1. Decision trees normally suffer from the problem of overfitting if it's allowed to grow without any control.	1. Random forests are created from subsets of data and the final output is based on average or majority ranking and hence the problem of overfitting is taken care of.
2. A single decision tree is faster in computation.	2. It is comparatively slower.
3. When a data set with features is taken as input by a decision tree it will formulate some set of rules to do prediction.	3. Random forest randomly selects observations, builds a decision tree and the average result is taken. It doesn't use any set of formulas.

Thus random forests are much more successful than decision trees only if the trees are diverse and acceptable.

Important Hyperparameters

Hyperparameters are used in random forests to either enhance the performance and predictive power of models or to make the model faster.

Following hyperparameters increases the predictive power:

1. **n_estimators**– number of trees the algorithm builds before averaging the predictions.
2. **max_features**– maximum number of features random forest considers splitting a node.
3. **mini_sample_leaf**– determines the minimum number of leaves required to split an internal node.

Following hyperparameters increases the speed:

4. **n_jobs**– it tells the engine how many processors it is allowed to use. If the value is 1, it can use only one processor but if the value is -1 there is no limit.
5. **random_state**– controls randomness of the sample. The model will always produce the same results if it has a definite value of random state and if it has been given the same hyperparameters and the same training data.
6. **oob_score** – OOB means out of the bag. It is a random forest cross-validation method. In this one-third of the sample is not used to train the data instead used to evaluate its performance. These samples are called out of bag samples.

Hardware requirements:

- System : Pentium i3 Processor.
- Hard Disk : 500 GB.
- Monitor : 15" LED
- Input Devices : Keyboard, Mouse

- Ram : 2 GB

Software requirements:

- Operating system : Windows 10.
- Coding Language : Python
- Tool: PYCHAM
- Libraries: Open CV, argparse, imutlis, math, dlib, pygam.

Chapter 4

4.1Machine learning

What are the 7 steps of machine learning?

7 Steps of Machine Learning

- Step 1: Gathering Data. ...
- Step 2: Preparing that Data. ...
- Step 3: Choosing a Model. ...
- Step 4: Training. ...
- Step 5: Evaluation. ...
- Step 6: Hyper parameter Tuning. ...
- Step 7: Prediction.

Introduction:

In this blog, we will discuss the workflow of a Machine learning project this includes all the steps required to build the proper machine learning project from scratch. We will also go over data pre-processing, data cleaning, feature exploration and feature engineering and show the impact that it has on Machine Learning Model Performance. We will also cover a couple of the pre-modelling steps that can help to improve the model performance.

Python Libraries that would be need to achieve the task:

1. Numpy
2. Pandas
3. Sci-kit Learn
4. Matplotlib

Understanding the machine learning workflow

We can define the machine learning workflow in 3 stages:

1. Gathering data

2. Data pre-processing
3. Researching the model that will be best for the type of data
4. Training and testing the model
5. Evaluation

What is the machine learning Model?

The machine learning model is nothing but a piece of code; an engineer or data scientist makes it smart through training with data. So, if you give garbage to the model, you will get garbage in return, i.e. the trained model will provide false or wrong prediction

1. Gathering Data

The process of gathering data depends on the type of project we desire to make, if we want to make an ML project that uses real-time data, then we can build an IoT system that using different sensors data. The data set can be collected from various sources such as a file, database, sensor and many other such sources but the collected data cannot be used directly for performing the analysis process as there might be a lot of missing data, extremely large values, unorganized text data or noisy data. Therefore, to solve this problem Data Preparation is done. We can also use some free data sets which are present on the internet. Kaggle and UCI Machine learning Repository are the repositories that are used the most for making Machine learning models. Kaggle is one of the most visited websites that is used for practicing machine learning algorithms, they also host competitions in which people can participate and get to test their knowledge of machine learning.

2. Data pre-processing

Data pre-processing is one of the most important steps in machine learning. It is the most important step that helps in building machine learning models more accurately. In machine learning, there is an 80/20 rule. Every data scientist should spend 80% time for data pre-processing and 20% time to actually perform the analysis.

What is data pre-processing?

Data pre-processing is a process of cleaning the raw data i.e. the data is collected in the real world and is converted to a clean data set. In other words, whenever the data is gathered from different sources it is collected in a raw format and this data isn't feasible for the analysis. Therefore, certain steps are executed to convert the data into a small clean data set, this part of the process is called as data pre-processing.

Why do we need it?

As we know that data pre-processing is a process of cleaning the raw data into clean data, so that can be used to train the model. So, we definitely need data pre-processing to achieve good results from the applied model in machine learning and deep learning projects. Most of the real-world data is messy, some of these types of data are:

1. **Missing data:** Missing data can be found when it is not continuously created or due to technical issues in the application (IOT system).
2. **Noisy data:** This type of data is also called outliers, this can occur due to human errors (human manually gathering the data) or some technical problem of the device at the time of collection of data.
3. **Inconsistent data:** This type of data might be collected due to human errors (mistakes with the name or values) or duplication of data.

Three Types of Data

1. Numeric e.g. income, age
2. Categorical e.g. gender, nationality
3. Ordinal e.g. low/medium/high

How can data pre-processing be performed?

These are some of the basic pre processing techniques that can be used to convert raw data.

1. **Conversion of data:** As we know that Machine Learning models can only handle numeric features, hence categorical and ordinal data must be somehow converted into numeric features.

2. **Ignoring the missing values:** Whenever we encounter missing data in the data set then we can remove the row or column of data depending on our need. This method is known to be efficient but it shouldn't be performed if there are a lot of missing values in the dataset.

3. **Filling the missing values:** Whenever we encounter missing data in the data set then we can fill the missing data manually, most commonly the mean, median or highest frequency value is used.

4. **Machine learning:** If we have some missing data then we can predict what data shall be present at the empty position by using the existing data.

5. **Outliers detection:** There are some error data that might be present in our data set that deviates drastically from other observations in a data set. [Example: human weight = 800 Kg; due to mistyping of extra 0]

Researching the model that will be best for the type of data

Our main goal is to train the best performing model possible, using the pre-processed data.

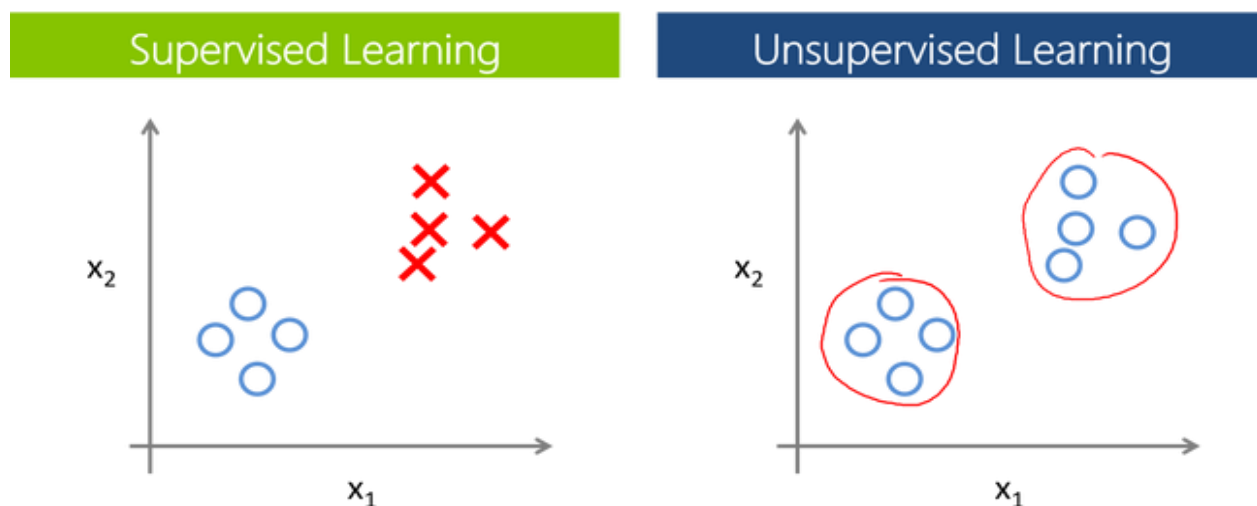


Fig 4.1 Model

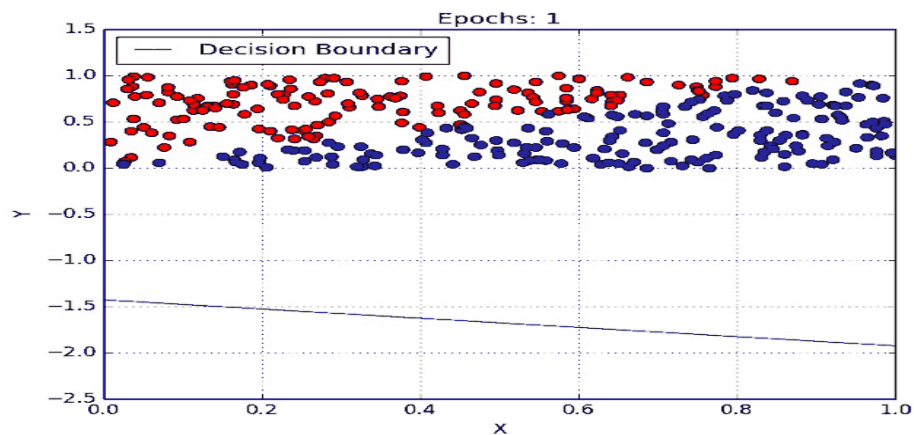
Supervised Learning:

In Supervised learning, an AI system is presented with data which is labelled, which means that each data tagged with the correct label.

The supervised learning is categorized into 2 other categories which are “**Classification**” and “**Regression**”.

Classification:

Classification problem is when the target variable is **categorical** (i.e. the output could be classified into classes — it belongs to either Class A or B or something else). A classification problem is when the output variable is a category, such as “red” or “blue” , “disease” or “no disease” or “spam” or “not spam”.



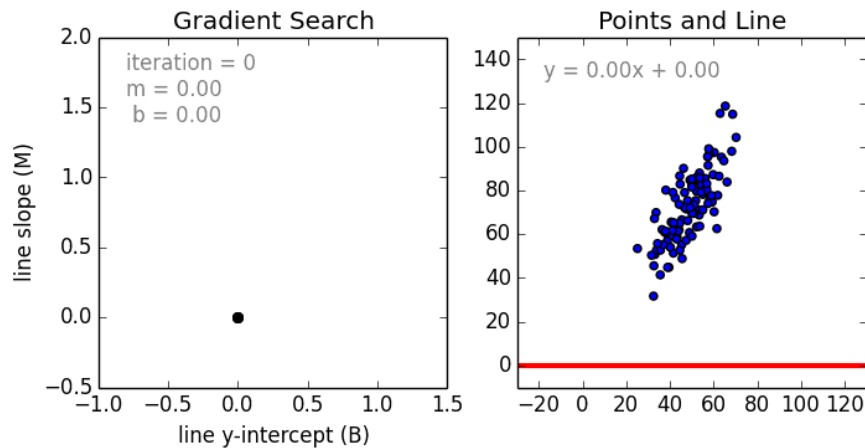
As shown in the above representation, we have 2 classes which are plotted on the graph i.e. red and blue which can be represented as 'setosa flower' and 'versicolor flower', we can image the X-axis as there 'Sepal Width' and the Y-axis as the 'Sepal Length', so we try to create the best fit line that separates both classes of flowers

These some most used classification algorithms.

- **K-Nearest Neighbor**
- **Naive Bayes**
- **Decision Trees/Random Forest**
- **Support Vector Machine**
- **Logistic Regression**

Regression:

While a Regression problem is when the target variable is continuous (i.e. the output is numeric).



As shown in the above representation, we can imagine that the graph's X-axis is the 'Test scores' and the Y-axis represents 'IQ'. So we try to create the best fit line in the given graph so that we can use that line to predict any approximate IQ that isn't present in the given data.

These some most used regression algorithms:

- **Linear Regression**
- **Support Vector Regression**
- **Decision Tress/Random Forest**
- **Gaussian Progresses Regression**
- **Ensemble Methods**

Unsupervised Learning:

The unsupervised learning is categorized into 2 other categories which are "Clustering" and "Association".

Clustering:

A set of inputs is to be divided into groups. Unlike in classification, the groups are not known beforehand, making this typically an unsupervised task.

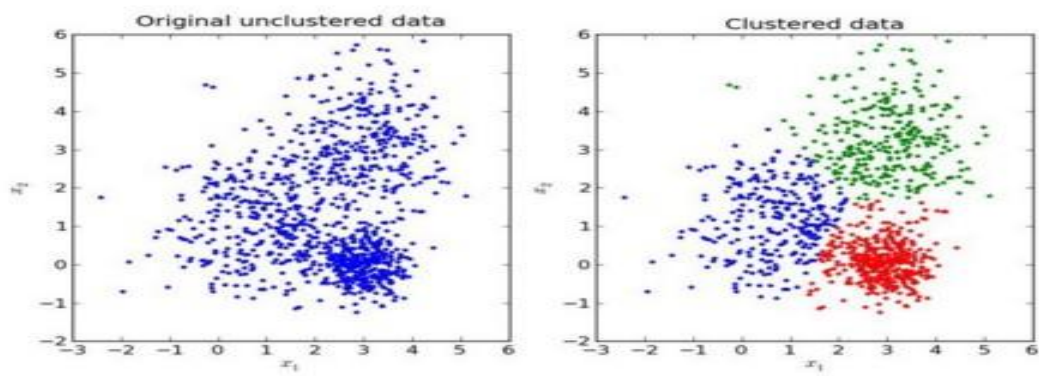


Fig 4.2 Clustering

Methods used for clustering are:

- **Gaussian mixtures**
- **K-Means Clustering**
- **Boosting**
- **Hierarchical Clustering**
- **K-Means Clustering**
- **Spectral Clustering**

Overview of models under categories:

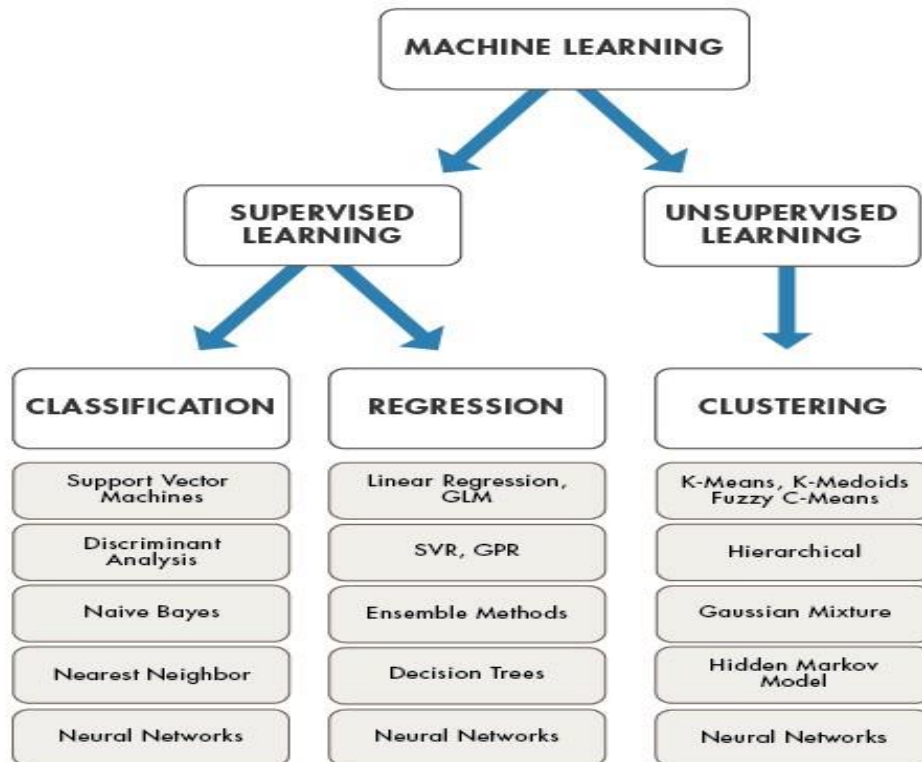


Fig 4.3 ML Overview

4. Training and testing the model on data

For training a model we initially split the model into 3 three sections which are 'Training data', 'Validation data' and 'Testing data'. You train the classifier using 'training data set', tune the parameters using 'validation set' and then test the performance of your classifier on unseen 'test data set'. An important point to note is that during training the classifier only the training and/or validation set is available. The test data set must not be used during training the classifier. The test set will only be available during testing the classifier.

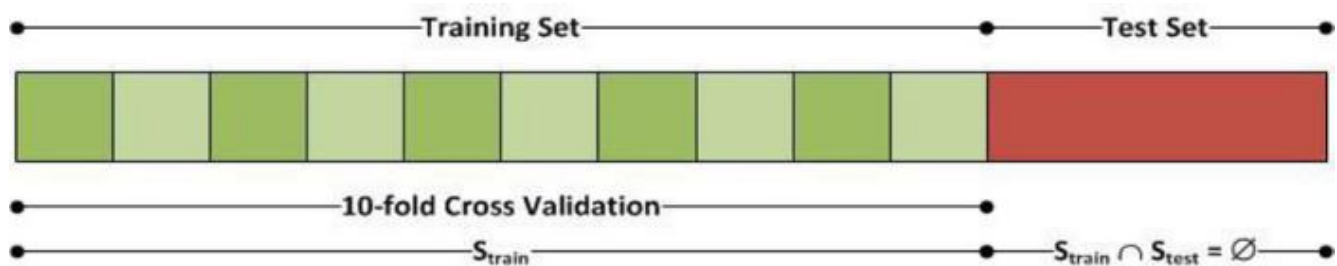


Fig 4.4 Training and Testing

Training set:

The training set is the material through which the computer learns how to process information. Machine learning uses algorithms to perform the training part. A set of data used for learning, that is to fit the parameters of the classifier.

Validation set:

Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. A set of unseen data is used from the training data to tune the parameters of a classifier.

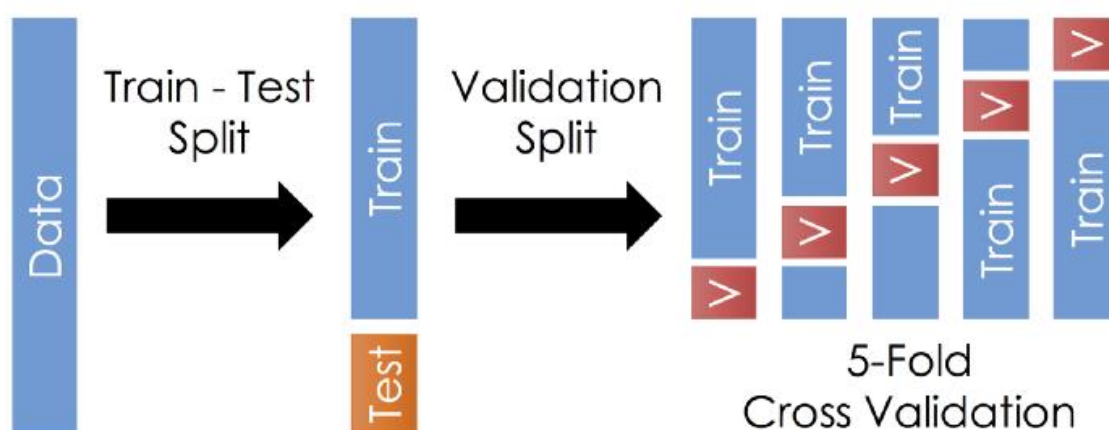


Fig 4.5 Validation Set

Once the data is divided into the 3 given segments we can start the training process.

In a data set, a training set is implemented to build up a model, while a test (or validation) set is to validate the model built. Data points in the training set are excluded from the test (validation) set. Usually, a data set is divided into a training set, a validation set (some people use 'test set' instead) in each iteration, or divided into a training set, a validation set and a test set in each iteration.

The model uses any one of the models that we had chosen in step 3/ point 3. Once the model is trained we can use the same trained model to predict using the testing data i.e. the unseen data. Once this is done we can develop a confusion matrix, this tells us how well our model is trained. A confusion matrix has 4 parameters, which are '**True positives**', '**True Negatives**', '**False Positives**' and '**False Negative**'. We prefer that we get more values in the True negatives and true

positives to get a more accurate model. The size of the Confusion matrix completely depends upon the number of classes.

n=165	Predicted: NO	Predicted: YES
	Actual: NO	Actual: YES
	50	10
	5	100

True positives : These are cases in which we predicted TRUE and our predicted output is correct.

True negatives : We predicted FALSE and our predicted output is correct.

False positives : We predicted TRUE, but the actual predicted output is FALSE.

False negatives : We predicted FALSE, but the actual predicted output is TRUE.

We can also find out the accuracy of the model using the confusion matrix.

Accuracy = (True Positives + True Negatives) / (Total number of classes)

i.e. for the above example:

Accuracy = (100 + 50) / 165 = 0.9090 (90.9% accuracy)

5.Evaluation

Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future.

To improve the model we might tune the hyper-parameters of the model and try to improve the accuracy and also looking at the confusion matrix to try to increase the number of true positives and true negatives.

6.Conclusion

In this blog, we have discussed the workflow a Machine learning project and gives us a basic idea of how a should the problem be tackled.

Chapter 5

SYSTEM DESIGN

Modules:

- Data Collection
- Per-Processing
- Feature Extraction
- Models and Training
- Testing Model
- Performance Evaluation
- Prediction

Data Collection :

Collecting data allows you to capture a record of past events so that we can use data analysis to find recurring patterns. From those patterns, you build predictive models using machine learning algorithms that look for trends and predict future changes. Predictive models are only as good as the data from which they are built, so good data collection practices are crucial to developing high-performing models. The data need to be error-free (garbage in, garbage out) and contain relevant information for the task at hand. For example, a loan default model would not benefit from tiger population sizes but could benefit from gas prices over time. In this module, we collect the data from kaggle dataset archives. This dataset contains the information of divorce in previous years.

Pre-Processing:

- Data cleaning is a critically important step in any machine learning project.
- In this module data cleaning is done to prepare the data for analysis by removing or modifying the data that may be incorrect, incomplete, duplicated or improperly formatted.

- In tabular data, there are many different statistical analysis and data visualization techniques you can use to explore your data in order to identify data cleaning operations you may want to perform.

Feature Extraction:

- This is done to reduce the number of attributes in the dataset hence providing advantages like speeding up the training and accuracy improvements.
- In machine learning, pattern recognition, and image processing, feature extraction starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps, and in some cases leading to better human interpretations. Feature extraction is related to dimensionality reduction
- When the input data to an algorithm is too large to be processed and it is suspected to be redundant (e.g. the same measurement in both feet and meters, or the repetitiveness of images presented as pixels), then it can be transformed into a reduced set of features (also named a feature vector).
- Determining a subset of the initial features is called feature selection. The selected features are expected to contain the relevant information from the input data, so that the desired task can be performed by using this reduced representation instead of the complete initial data.

Model training:

- A training model is a dataset that is used to train an ML algorithm. It consists of the sample output data and the corresponding sets of input data that have an influence on the output.
- The training model is used to run the input data through the algorithm to correlate the processed output against the sample output. The result from this correlation is used to modify the model.
- This iterative process is called “model fitting”. The accuracy of the training dataset or the validation dataset is critical for the precision of the model.
- Model training in machine language is the process of feeding an ML algorithm with data to help identify and learn good values for all attributes involved.
- There are several types of machine learning models, of which the most common ones are supervised and unsupervised learning.

- In this module we use supervised classification algorithms like linear regression to train the model on the cleaned dataset after dimensionality reduction.

Testing model:

- In this module we test the trained machine learning model using the test dataset
- Quality assurance is required to make sure that the software system works according to the requirements. Were all the features implemented as agreed? Does the program behave as expected? All the parameters that you test the program against should be stated in the technical specification document.
- Moreover, software testing has the power to point out all the defects and flaws during development. You don't want your clients to encounter bugs after the software is released and come to you waving their fists. Different kinds of testing allow us to catch bugs that are visible only during runtime.

Performance Evaluation:

- In this module, we evaluate the performance of trained machine learning model using performance evaluation criteria such as F1 score, accuracy and classification error.
- In case the model performs poorly, we optimize the machine learning algorithms to improve the performance.
- Performance Evaluation is defined as a formal and productive procedure to measure an employee's work and results based on their job responsibilities. It is used to gauge the amount of value added by an employee in terms of increased business revenue, in comparison to industry standards and overall employee return on investment (ROI).
- All organizations that have learned the art of "winning from within" by focusing inward towards their employees, rely on a systematic performance evaluation process to measure and evaluate employee performance regularly.
- Ideally, employees are graded annually on their work anniversaries based on which they are either promoted or are given suitable distribution of salary raises
- Performance evaluation also plays a direct role in providing periodic feedback to employees, such that they are more self-aware in terms of their performance metrics.

Prediction:

- Prediction refers to the output of an algorithm after it has been trained on a historical dataset and applied to new data when forecasting the likelihood of a particular outcome, such as whether or not a customer will churn in 30 days.
- The algorithm will generate probable values for an unknown variable for each record in the new data, allowing the model builder to identify what that value will most likely be.
- The word “prediction” can be misleading. In some cases, it really does mean that you are predicting a future outcome, such as when you’re using machine learning to determine the next best action in a marketing campaign.
- Other times, though, the “prediction” has to do with, for example, whether or not a transaction that already occurred was fraudulent.
- In that case, the transaction already happened, but you’re making an educated guess about whether or not it was legitimate, allowing you to take the appropriate action.

Chapter 6

Conclusions:

- The project has successfully met its aim and is ready to predict all the future behavior and trends of the market based on the then scenario. By this prediction model one can easily find out the possibility of the growth and sales of the company and can take actions accordingly. This model is ready to serve the purpose of fulfilling the prediction of sales and production after the impact of the corona virus outbreak.
- Further if anything this sort or any other pandemic hit again then this model is going to suggest how to save the companies worth by providing the required details through its predictions. The model has been currently at a base level and covers few factors currently so it depends upon many circumstances and conditions.
- Further it has a lot of area to be worked upon and lots of factors to be taken care of and considered to give more accurate results. The basic purpose of making this model is to save companies from pandemic such as COVID-19. They can be insured with this type of a model and takes the necessary steps according to the benefit of the company in the upcoming time.
- The model has been predicting the possibility of saving the worth of a company and preventing them economically by not misusing their wealth. It gives a financial preparation to the company by predicting the future outcomes. The automobile industry has been in a very delicate situation, they need to listen to their customers to move forward treading carefully. Listening to the demand has

been the only viable option for them to create sales in such times. If they stay patient then surely the tides will turn and the situation will get better, it has been just a matter of sticking to the basics.

REFERENCES:

- Friedman, J.H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* 2020, 38, 367–378. [CrossRef]
- Pavlyshenko, B.M. Linear, machine learning and probabilistic approaches for time series analysis. In *Proceedings of the IEEE First International Conference on Data Stream Mining & Processing (DSMP)*, Lviv, Ukraine, 23–27 August 2016; IEEE: Piscataway, NJ, USA, 2020; pp. 377–381.
- Pavlyshenko, B. Machine learning, linear and Bayesian models for logistic regression in failure detection problems. In *Proceedings of the 2016 IEEE International Conference on Big Data (Big Data)*, Washington, DC, USA, 5–8 December 2020; IEEE: Piscataway, NJ, USA
- Pavlyshenko, B. Using Stacking Approaches for Machine Learning Models. In *Proceedings of the 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)*, Lviv, Ukraine, 21–25 August 2018; IEEE: Piscataway, NJ, USA, 2018;
- Akkarapol Sa-ngasoongsong, Satish T.S. Bukkapatnam, Jaebeom Kim, Parameshwaran S. Iyer and R.P. Sureshc, "Multi -step sales forecasting in automotive industry based on structural relationship identification", *International Journal of Production Economics*, vol. 140, no. 2, pp. 875-887, December 2020.

Requirements

Flask==1.1.2

numpy==1.20.1

matplotlib==3.3.4

pandas==1.2.4

scikit_learn==0.2

APPENDICES

A. SOURCE CODE:

```
import numpy as np

from flask import Flask, request, jsonify, render_template

import pickle

import json

app = Flask(__name__)

model = pickle.load(open('model.pkl','rb'))

myjson=open('columns.json','r')

jsonData=myjson.read()

obj=json.loads(jsonData)

r=0

for x in obj:

    if isinstance(obj[x], list):

        r += len(obj[x])

def predict_price(year, kilometers_driven, mileage, engine, power, seats,
company, name, location, fuel_type, transmission, owner_type):

    x2 = [0]*r
```

```

x2[0] = int(year)

x2[1] = kilometers_driven

x2[2] = mileage

x2[3] = engine

x2[4] = power

x2[5] = seats

l=[]

for w in obj:

    for y in obj[w]:

        l.append(y)

w=l

for i in range(6,r):

    if w[i]==company:

        x2[i] = 1

    if w[i]==name:

        x2[i] = 1

    if w[i]==location:

        x2[i] = 1

    if w[i]==fuel_type:

        x2[i] = 1

```

```

        if w[i]==transmission:

            x2[i] = 1

        if w[i]==owner_type:

            x2[i] = 1

    result = ([x2])

    return result


def xyz(ash):

    c=predict_price(ash[0],ash[1],ash[2],ash[3],ash[4],ash[6],ash[5],ash[7],ash[8],ash[9]
    ],ash[10],ash[11])

    return c


@app.route('/')

def home():

    return render_template('index.html')


@app.route('/predict',methods=['POST'])

def predict():

    int_features = [x for x in request.form.values()]

```

```
final_features = list(int_features)
```

```
k=xyz(final_features)
```

```
prediction = model.predict(k)
```

```
output = prediction[0]
```

```
return render_template('index.html',prediction_text='Price is ₹ {}  
lakhs'.format(output))
```

```
if __name__ == "__main__":
```

```
    app.run(debug=True)
```

its app.py

```
import numpy as np
```

```
from sklearn import metrics
```

```
from sklearn.metrics import confusion_matrix, r2_score
```

```
from sklearn.model_selection import cross_val_score
```

```
from sklearn.model_selection import ShuffleSplit
```

```
import pandas as pd
```

```
import matplotlib.pyplot as plt
```

```
import matplotlib

matplotlib.rcParams["figure.figsize"] = (20,10)

import warnings

warnings.filterwarnings('ignore')

import seaborn as sns


# ### Data Load: Load used car prices into a dataframe


df1 = pd.read_csv('train-data.csv')

df1.head()


df1.shape


df1.columns


df1.isnull().sum()

cols_to_log = df1.select_dtypes(include=np.number).columns.tolist()

for colname in cols_to_log:

    sns.distplot(df1[colname], kde=True)
```

```
plt.show()

# # Data Cleaning: Handle NA values

# ##### Drop features that are containing 70% data null

# In our dataset we had only New_Price which contain above 70% data

df2 = df1.dropna(thresh=df1.shape[0]*0.7, how='all',axis=1)


df2.head(2)

df2.isnull().sum()

df2.shape

df3 = df2.dropna()

df3.isnull().sum()

df3.shape

# # Feature Engineering: Handle Categorical Variable

# ##### Create a company using Name column
```

```
company = [i.split()[0] for i in df3['Name']]
```

```
df3.insert(0, "Company", company)
```

```
# ##### Converting string data to numeric form, i.e removing units etc.
```

```
df3['Mileage'] = pd.to_numeric(df3['Mileage'].str.lower().str.split().str.get(0),  
errors='coerce')
```

```
df3['Engine'] = pd.to_numeric(df3['Engine'].str.lower().str.split().str.get(0),  
errors='coerce')
```

```
df3['Power'] = pd.to_numeric(df3['Power'].str.lower().str.split().str.get(0),  
errors='coerce')
```

```
# # Data Analysis with Visualisation
```

```
df3['Company'].value_counts()
```

```
company_count=df3['Name'].str.lower().str.split().str.get(0).to_frame()["Name"].value_counts()
```

```
plt.figure(figsize=(12,12))
```

```
plt.xlabel('Company Name')

plt.ylabel('No. of cars sold')

plt.title('Cars sold per company')

company_count.plot(kind='bar')

plt.show()
```

```
mileage = df3.groupby('Fuel_Type').Mileage.mean()

plt.xlabel("Fuel Type")

plt.xlabel("Mileage")

plt.title("Fuel Type vs Mileage")

mileage.plot(kind='bar')

plt.show()
```

```
purchased_car_per_year = df3['Year'].value_counts()

purchased_car_per_year.plot(kind='bar')

plt.xlabel("Year")

plt.ylabel("Paurchased Cars")

plt.title("Year vs Purchased cars")
```

```
plt.show()
```

```
average_price = df3.groupby('Company').Price.mean()
```

```
average_price.plot(kind='bar')
```

```
plt.show()
```

```
Company_Kilometers_Driven =
```

```
df3.groupby('Company').Kilometers_Driven.mean()
```

```
Company_Kilometers_Driven.plot(kind='bar')
```

```
plt.xlabel("s")
```

```
plt.ylabel("s")
```

```
plt.title("Average Kilometeres vs Company")
```

```
plt.show()
```

```
df3.head(2)
```

```
df3.Name = df3.Name.apply(lambda x: x.strip())
```

```
name_stats = df3['Name'].value_counts(ascending=False)
```

```
name_stats
```

```
name_stats.head()
```

```
name_stats.values.sum()
```

```
len(name_stats[name_stats>10])
```

```
len(name_stats)
```

```
len(name_stats[name_stats<=10])
```

```
name_stats_less_than_10 = name_stats[name_stats<=10]
```

```
name_stats_less_than_10
```

```
len(df3.Name.unique())
```

```
df3.Name
```

```
df3.Name = df3.Name.apply(lambda x: 'other' if x in name_stats_less_than_10  
else x)
```

```
len(df3.Name.unique())
```

```
df3.head(10)
```

```
df3[df3.Mileage<5].head()
```

```
mileage_mean = df3.Mileage.mean()
```

```
df3['Mileage'] = [i if i > 5 else float(int(mileage_mean)) for i in df3.Mileage]
```

```
df3.describe()
```

```
# Check Outliers in Kilometers_Driven
```

```
df3['Kilometers_Driven'].min()
```

```
df3['Kilometers_Driven'].max()
```

```
df3.sort_values(by='Kilometers_Driven', ascending=False)
```

```
# ##### Check above data points. We have 65Million Kilometers_Driven and It is  
not possible. So we will handle this outlier using mean of Kilometer_Driven
```

```
kilometers_driven_mean = df3['Kilometers_Driven'].mean()
```

```
df3['Kilometers_Driven'] = [i if i < 1000000 else float(int(kilometers_driven_mean))  
for i in df3.Kilometers_Driven]
```

```
# Maximum Driven
```

```
df3['Kilometers_Driven'].max()
```

```
df3.isnull().sum()
```

```
df3['Power'] = df3['Power'].fillna(df3['Power'].mean())
```

```
print(df3['Company'].value_counts())
```

```
print(df3['Location'].value_counts())
```

```
print(df3['Fuel_Type'].value_counts())
```

```
print(df3['Transmission'].value_counts())
```

```
print(df3['Owner_Type'].value_counts())
```

```
# Convert year variable in age
```

```
df3['Year'] = 2021-df3['Year']
```

```
df3.head()
```

```
df3.head(2)
```

```
# In[53]:
```

```
company_dummies = pd.get_dummies(df3['Company'])
```

```
name_dummies = pd.get_dummies(df3['Name'])
```

```
location_dummies = pd.get_dummies(df3['Location'])
```

```
fuel_type_dummies = pd.get_dummies(df3['Fuel_Type'])
```

```
transmission_dummies = pd.get_dummies(df3['Transmission'])
```

```
owner_type_dummies = pd.get_dummies(df3['Owner_Type'])
```

```
# In[54]:
```

```
df3.head()
```

```
# Create new dataframe with numeric variables
```

```
features = df3[['Year', 'Kilometers_Driven', 'Mileage', 'Engine', 'Power', 'Seats',  
'Price']]
```

```
df4 = pd.concat([features, company_dummies, name_dummies,  
location_dummies, fuel_type_dummies, transmission_dummies,  
owner_type_dummies], axis=1)
```

```
df4.shape
```

```
df4.head()
```

```
# ## Build Model Now...
```

```
# In[59]:
```

```
x = df4.drop(['Price'], axis=1)
```

```
y = df4['Price']
```

```
# In[60]:
```

```
x.shape
```

```
sns.heatmap(df1.corr(), annot=True, cmap='Blues',linewidth=0.2)
```

```
fig = plt.gcf()
```

```
fig.set_size_inches(10,8)
```

```
plt.show()
```

```
# In[61]:
```

```
from sklearn.preprocessing import MinMaxScaler
```

```
x[['Kilometers_Driven', 'Mileage']] =
```

```
MinMaxScaler().fit_transform(x[['Kilometers_Driven', 'Mileage']])
```

```
# In[62]:
```

```
from sklearn.model_selection import train_test_split

x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2,
random_state=0)
```

```
# In[63]:
```

```
from sklearn.neighbors import KNeighborsRegressor

knn = KNeighborsRegressor(n_neighbors=2)

knn.fit(x_train, y_train)

y_pred = knn.predict(x_test)

r2 = r2_score(y_test, y_pred)

print(f'R2 SCORE of KNN: {(r2_score(y_test, y_pred)*100)}')
```

```
fig, ax = plt.subplots()
```

```

ax.scatter(y_test, y_pred)

ax.plot([y.min(), y.max()], [y.min(), y.max()], 'k--', lw=4)

ax.set_xlabel('Actual')

ax.set_ylabel('Predicted')

ax.set_title('R2 score of KNN')


plt.show()


from sklearn.ensemble import RandomForestRegressor

rf = RandomForestRegressor(n_estimators = 1)

rf.fit(x_train, y_train)

y_pred1 = rf.predict(x_test)

r2_score(y_test, y_pred1)

print(f'R2 SCORE of RandomForest: {(r2_score(y_test, y_pred1)*100)}')


# In[64]:


df4.head()

```

```
cv = ShuffleSplit(n_splits=5, test_size=0.2, random_state=0)
```

```
cross_val_score(KNeighborsRegressor(), x, y, cv=cv)
```

```
x.head()
```

```
import pickle
```

```
pickle.dump(knn, open('model.pkl','wb'))
```

```
# In[88]:
```

```
import json
```

```
columns = {
```

```
    'data_columns' : [col.lower() for col in x.columns]
```

```
}
```

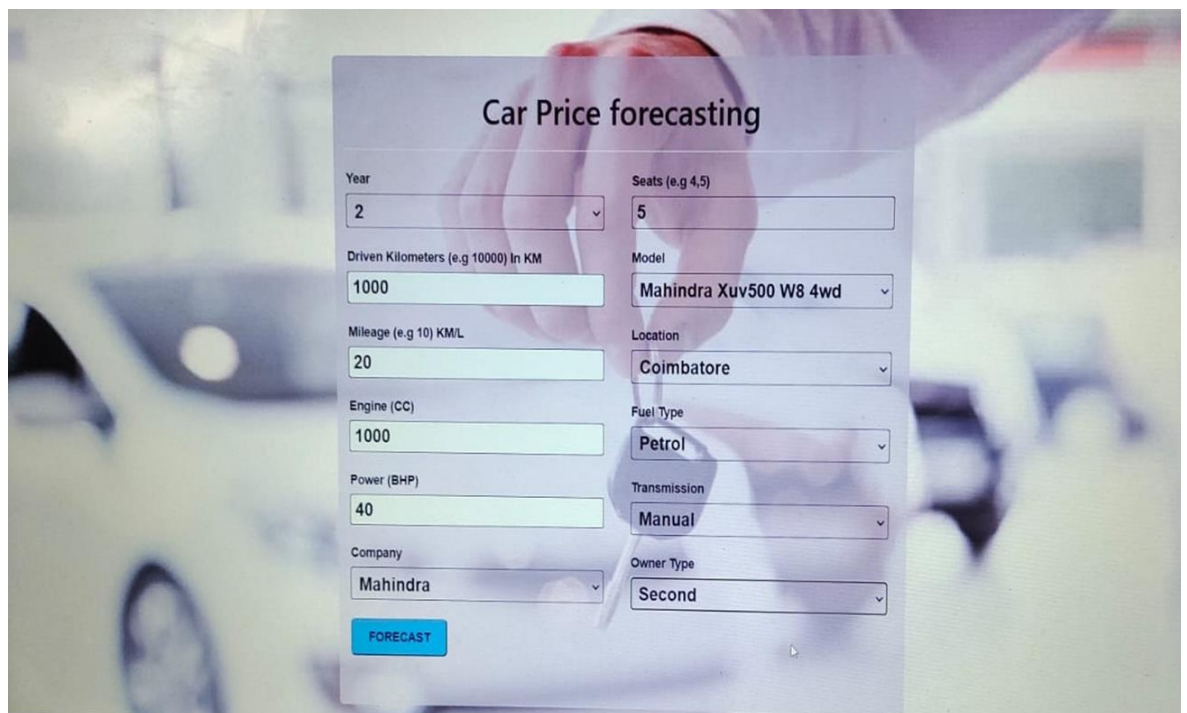
with open("columns.json","w") as f:

```
f.write(json.dumps(columns))
```

model.py

B. SCREENSHOTS:-

Fig 1: Parameters Entry



The screenshot shows a web application titled "Car Price forecasting". It features a form with the following fields and values:

Field	Value
Year	2
Seats (e.g 4,5)	5
Driven Kilometers (e.g 10000) In KM	1000
Model	Mahindra Xuv500 W8 4wd
Mileage (e.g 10) KM/L	20
Location	Coimbatore
Engine (CC)	1000
Fuel Type	Petrol
Power (BHP)	40
Transmission	Manual
Company	Mahindra
Owner Type	Second

A blue button labeled "FORECAST" is located at the bottom left of the form.

FIG 2 RESULT

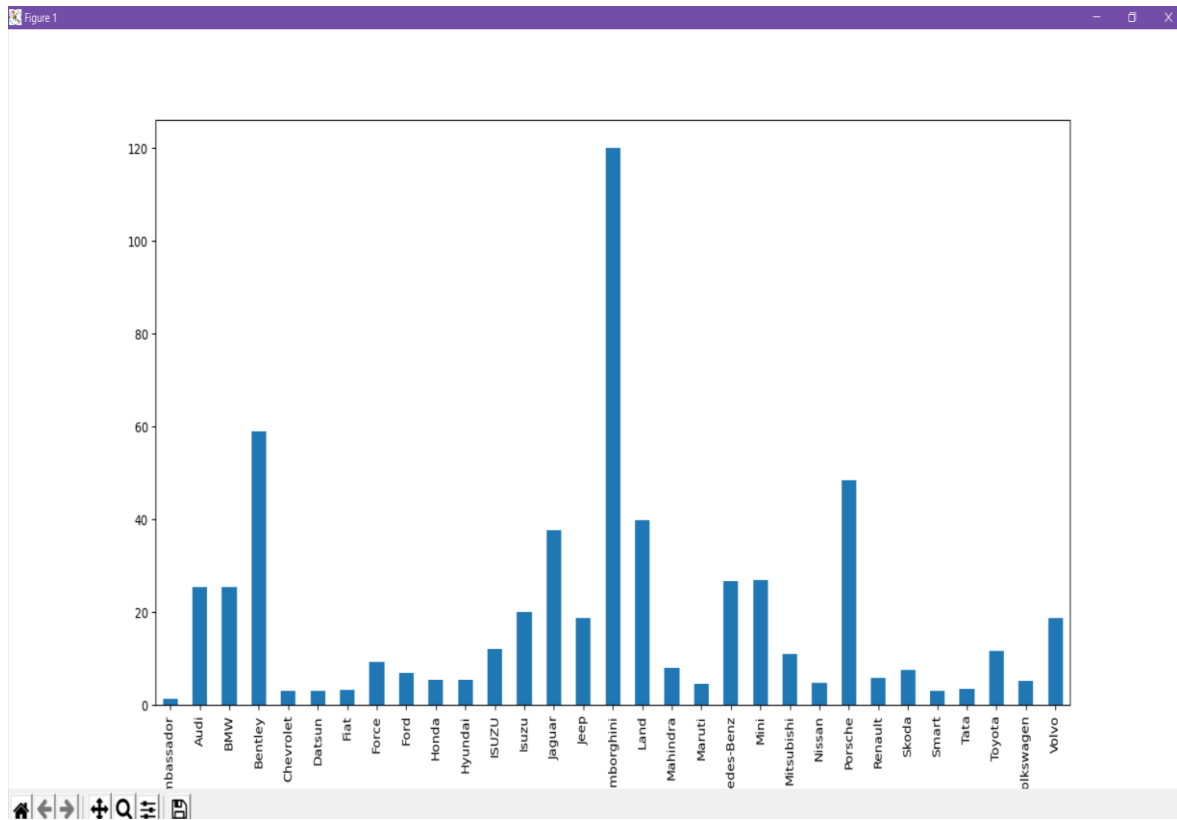
Car Price forecasting

Year 2	Seats (e.g 4,5)
Driven Kilometers (e.g 10000) In KM 	Model Volkswagen Polo Diesel Coml
Mileage (e.g 10) KM/L 	Location Coimbatore
Engine (CC) 	Fuel Type Lpg
Power (BHP) 	Transmission Manual
Company Bmw	Owner Type First

FORECAST

Price is ₹ 3.275 lakhs

FIG:2 Result



CORRELATIONAL MATRIX

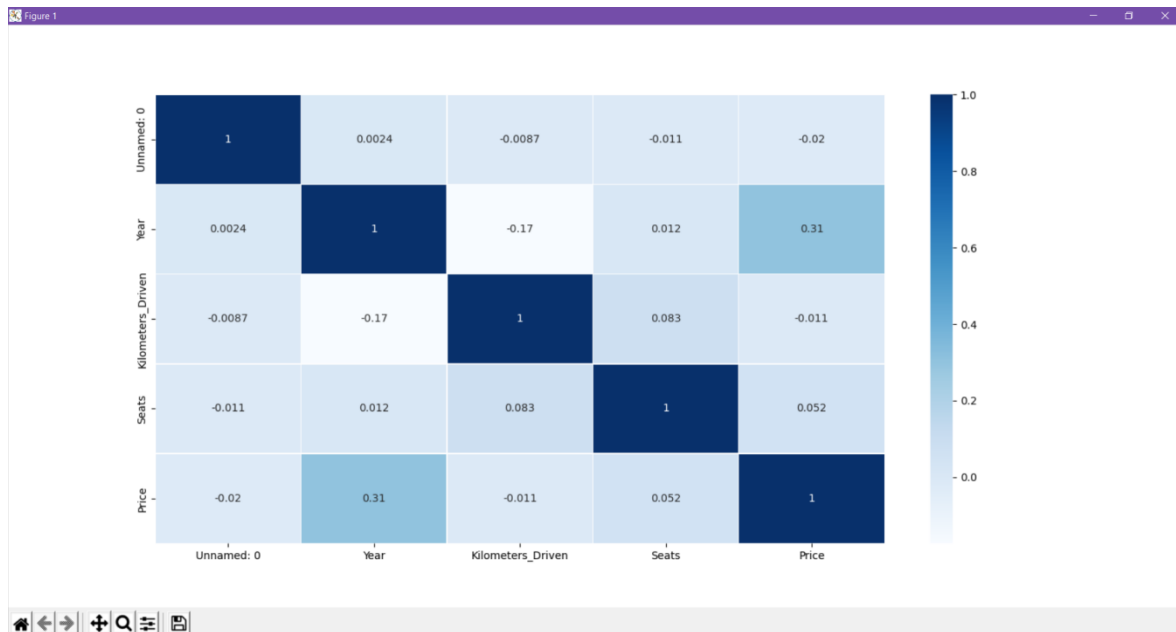


Fig 3 Feature Importance

C. PLAGIARISM REPORT :

SALES FORECAST OF MANUFACTURING COMPANIES USING MACHINE LEARNING NAVIGATING THE PANDEMIC LIKE COVID.docx

ORIGINALITY REPORT

1 %

SIMILARITY INDEX

1 %

INTERNET SOURCES

1 %

PUBLICATIONS

1 %

STUDENT PAPERS

PRIMARY SOURCES

1

ieeexplore.ieee.org

Internet Source

<1 %

2

onlinelibrary.wiley.com

Internet Source

<1 %

D. JOURNAL PAPER :

SALES FORECAST OF MANUFACTURING COMPANIES USING MACHINE LEARNING NAVIGATING THE PANDEMIC LIKE COVID-19

Venkata Manas
Cheruku
Student
Department of
CSE
Sathyabama
Institute of
Science and
Technology
Chennai, India

Venkatamanas1@
gmail.com

Kakumanu Eswar
Student
Department of
CSE
Sathyabama
Institute of

Science and
Technology
Chennai, India
eswarkakumanu
@gmail.com

Dr. S. Murugan
Associate
Professor

Department of
CSE
Sathyabama
Institute of
Science and
Technology
Chennai, India
maheswari.cse@s
athyabama.ac.in

REFERENCES

- Friedman, J.H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* 2020, 38, 367–378. [CrossRef]
- Pavlyshenko, B.M. Linear, machine learning and probabilistic approaches for time series analysis. In *Proceedings of the IEEE First International Conference on Data Stream Mining & Processing (DSMP)*, Lviv, Ukraine, 23–27 August 2016; IEEE: Piscataway, NJ, USA, 2020; pp. 377–381.
- Pavlyshenko, B. Machine learning, linear and Bayesian models for logistic regression in failure detection problems. In *Proceedings of the 2016 IEEE International Conference on Big Data (Big Data)*, Washington, DC, USA, 5–8 December 2020; IEEE: Piscataway, NJ, USA
- Pavlyshenko, B. Using Stacking Approaches for Machine Learning Models. In *Proceedings of the 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)*, Lviv, Ukraine, 21–25 August 2018; IEEE: Piscataway, NJ, USA, 2018;
- Akkarapol Sa-ngasoongsong, Satish T.S. Bukkapatnam, Jaebeom Kim, Parameshwaran S. Iyer and R.P. Sureshc, "Multi -step sales forecasting in automotive industry based on structural relationship identification", *International Journal of Production Economics*, vol. 140, no. 2, pp. 875-887, December 2020.

Biography

Master Venkata Manas Cheruku graduating from the beloved university Sathyabama Institute of Science and Technology in the stream of Computer Science and Engineering has excelled in this stream. His research interest is Machine Learning, Data Science and Cyber Security.

Master Kakumanu Eswar graduating from the beloved university Sathyabama Institute of Science and Technology in the stream of Computer Science and Engineering has excelled in this stream. His research interest is Machine Learning, Neural Network and Web Development.

Dr. S. Murugan, Associate professor in the Department of Computer Science and Engineering at Sathyabama Institute of Science and Technology in

Chennai. Her area of specialization is Database Management Systems, Data Mining, Machine Learning, and Data Analytics. She has authored in various reputed conferences and in international journals.