NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY

(AN AUTONOMOUS INSTITUTION, AFFILIATED TO VISVESVARAYA TECHNOLOGICAL UNIVERSITY, BELGAUM, APPROVED BY AICTE & GOVT.OF KARNATAKA

COURSE REPORT

ON

Introduction to Machine Learning

Submitted in partial fulfillment of the requirement for the award of Degree of

Bachelor of Engineering

in

Computer Science and Engineering

Submitted by:

Shreya Roy	1NT18CS155
Soumya Ramesh	1NT18CS216

Department of Computer Science and Engineering

2021-2022

Nitte Meenakshi Institute of Technology

(AN AUTONOMOUS INSTITUTION AFFILIATED TO VISVESVARAYA TECHNOLOGICAL UNIVERSITY,
BELGAUM)

PB No. 6429, Yelahanka, Bangalore 560-064, Karnataka

Telephone: 080- 22167800, 22167860

Fax: 080 - 22167805

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the **Course Repor**t titled *"Used Car Price Prediction*" is an authentic work carried out by **Shreya Roy(1NT18CS155)** and **Soumya Ramesh (1NT18CS216)** bonafide students of Nitte Meenakshi Institute of Technology, Bangalore in partial fulfillment for the award of the degree of Bachelor of Engineering in **COMPUTER SCIENCE AND ENGINEERING** of Visvesvaraya Technological University, Belagavi during the academic year 2021-2022.

Name Signature of the Faculty Incharge

Name and Signature of the HOD

Dr. Saroja Devi H

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompany the successful completion of any task would be incomplete without the mention of the people who made it possible, whose constant guidance and encouragement crowned our effort with success. We express our sincere gratitude to our Principal **Dr. H. C. Nagaraj**, Nitte Meenakshi Institute of Technology for providing facilities.

We wish to thank our HoD, **Dr. Saroja Devi H** for the excellent environment created to further educational growth in our college. We also thank him for the invaluable guidance provided which has helped in the creation of a better technical report.

Thanks to our Subject Faculty. We also thank all our friends, teaching and non-teaching staff at NMIT, Bangalore, for all the direct and indirect help provided in the completion of the presentation.

Lastly, Thanks to our parents for always supporting, encouraging and loving us.

Signature

Name	USN	
Shreya Roy	1NT18CS155	Shreya Roy
Soumya Ramesh	1NT18CS216	

Date: 20/12/2021

ABSTRACT

The pre-owned car market is expected to outpace the new car market, with the industry expected to clock a healthy growth rate of 15 per cent in FY22. 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. In this project we are using machine learning models such as naive bayesian classifier, linear regression model and decision tree model. The dataset is taken from Kaggle and is collected by Avi Kasliwal. It has more than 6000 rows of data. For validation, we will be using K-Fold cross validation and Train-Test Split validation. The results of the project will include visual graphs and plots of the testing data, identifying the relations and dependencies of the different attributes of the dataset, all the plots related to the machine learning algorithms being used, predicted price attribute csv file, plots of any observed characteristics or trends, from the result.

TABLE OF CONTENTS

INTRODUCTION

According to a study, the Indian used car market would increase to more than 70 lakh vehicles by 2025-26, up from 38 lakh in 2020-21, and will grow at a rate of 12-14 percent in the next few years. The pre-owned automobile sector is expected to grow at a 15% annual rate in FY22, according to the OLX Autos-CRISIL Study 2021, with the COVID-19 epidemic, digitalization, changing demographics and ambitions, first-time customers, and the availability of financing options serving as growth drivers. According to the study, many customers had previously avoided purchasing personal vehicles, particularly in metro areas, due to the growth of shared mobility providers such as OLA and Uber. The pandemic has had an effect on the market and customer behavior, reigniting enthusiasm in owning a personal automobile[1].

It's not easy to predict the selling price of a used car. It requires extensive domain knowledge as well as estimates of car features. Some of the important elements we need to consider when determining the price of used cars include the brand model, age of the car, color of the car, kilometers driven, owner and seller type, and fuel type of the car. The list of attributes is not restricted; we can explore a wide range of options that can improve the accuracy of our forecasts[2].

We are going to use linear regression model, random forest regressor and decision tree to predict the price of the used car based on its features. Linear regression is one of the most straightforward and widely used Machine Learning techniques. It's a statistical strategy for predicting outcomes. Linear regression makes predictions for variables that are either continuous/real or numeric[4].

Random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned.[5].

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome[8].

DATASET

We got the data from Kaggle, which is an open source and community-maintained website that was gathered by Avi Kasliwal. The dataset is called ‘Used Car Price Prediction’.

The train-data.csv contains the training data. It has 6019 rows and 14 columns in the training data. The dataset [3] contains both numerical and categorical data, with numerical data in the selling price and kilometers driven columns and categorical data in the remaining input fields. The attributes include:

1. Index: It stores data from 0 to 6018.
2. Name: The brand and model of the car. It includes over 1400 unique car names or models from well-known 29 car brands such as Maruti, Hyundai, Datsun, Jeep, BMW, Mahindra, Ford, Nissan, Tata, Chevrolet, Toyota, Jaguar, Mercedes-Benz, Audi, Skoda, Volvo, MG, Force, Isuzu, Opel Corsa, Ambassador, Kia, Renault, Fiat, Volkswagen, and many others.
3. Location: It denotes the year in which the car is being sold or is available for purchase. Our database includes a variety of automobiles purchased between 2001and 2019.
4. Year: The year or edition of the model.
5. Kilometers_Driven: It reflects the total number of kilometers driven by a vehicle. Our data includes a variety of vehicles that have traveled between 1 and 800,000 kilometers.
6. Fuel_Type: This reflects the type of fuel used by the car. It will be one among Petrol, Diesel, Electric, CNG, LPG.
7. Transmission: This represents the type of transmission used by the car. It will be one among Automatic or Manual.
8. Owner_type: Whether the ownership is Firsthand, Second hand or other.
9. Mileage: The standard mileage offered by the car company in kmpl or km/kg
10. Engine: The displacement volume of the engine in cc.
11. Power: The maximum power of the engine in bhp.
12. Seats: The number of seats in the car.
13. New_Price: The price of a new car of the same model.
14. Price: The price of the used car in INR Lakhs.

test-data.csv contains the test data. There are 1234 rows and 13 columns in the test data. It contains the first 13 columns. We need to predict the price of the used car after training the model.

The dataset selected by us showed inconsistencies in terms of the entries having null values. Also the Dataset has units as strings, which causes irregularities while computing if not corrected. Due to the huge size of the datasets, they have to be cleared of the duplicate and unnecessary values which have no effect on our model computation. The model name and Brand name of cars have to be considered separately so as to make our dataset less complex to process.

MACHINE LEARNING METHODS

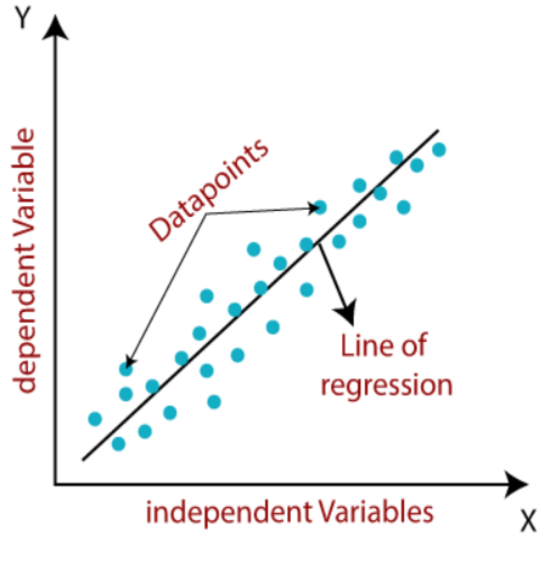
The project deals with , implementing and analyzing two Machine Learning Algorithms , namely the Linear Regression model, the Naive Bayes Classifier and the Decision Tree algorithm , thus driving a comparison between them to see which gives the best result in terms of accuracy and least amount of error.

The goal behind making use of these models is to resolve the problem , by predicting the price of Used Cars , as closely and accurately in real time as possible.

Following is brief explanation about the Machine Learning Models that will be used in the project:

1. **Linear Regression Model :**

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as **sales, salary, age, product price,** etc. Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable[4].



Mathematically, we can represent a linear regression:

$$y = a_0 + a_1 x + \varepsilon$$

Here,

Y= Dependent Variable (Target Variable)

X= Independent Variable (predictor Variable)

a0= intercept of the line (Gives an additional degree of freedom)

a1 = Linear regression coefficient (scale factor to each input value).

ε = random error

Fig1: Linear Regression

The values for x and y variables are training datasets for Linear Regression model representation[4].

2. Random Forest Regression:

Random Forest Regression is a supervised learning algorithm that uses ensemble learning methods for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.[6]

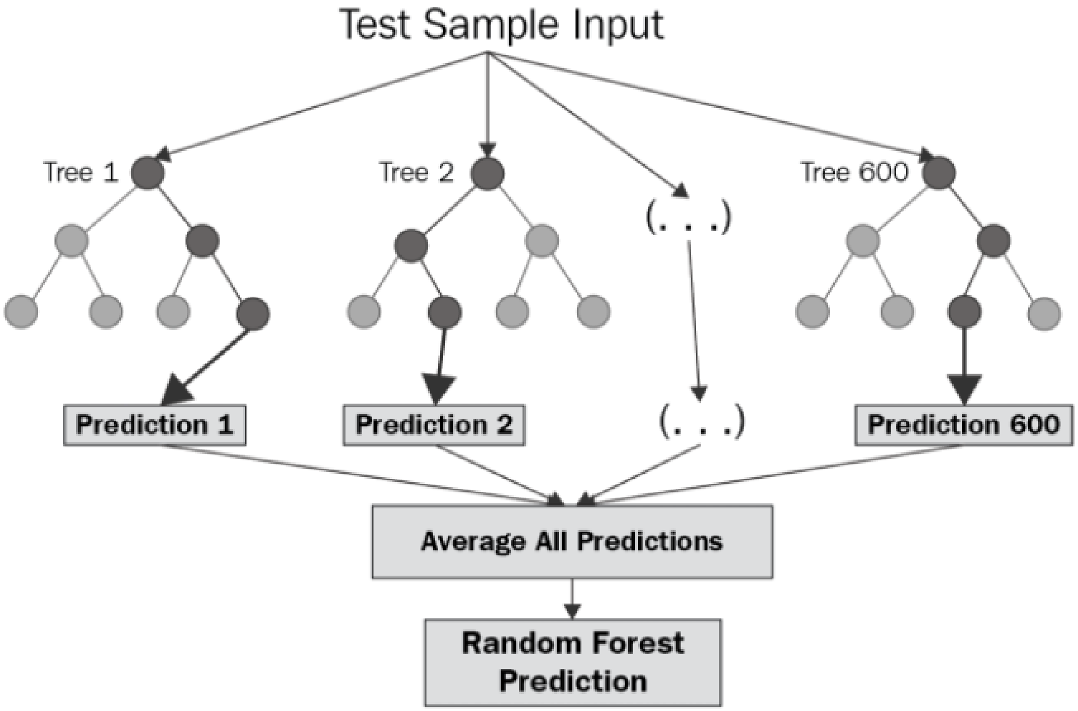


Fig 2. Random Forest Tree

The steps of the algorithm include:

- Pick at random k data points from the training set.
- Build a decision tree associated with these k data points.
- Choose the number N of trees you want to build and repeat steps 1 and 2.
- For a new data point, make each one of your N-tree trees predict the value of y for the data point in question and assign the new data point to the average across all of the predicted y values.[6]

The advantages of using this method is:

- It reduces overfitting in decision trees and helps to improve the accuracy

- It is flexible to both classification and regression problems
- It works well with both categorical and continuous values
- It automates missing values present in the data
- Normalizing of data is not required as it uses a rule-based approach.[7]

However, despite these advantages, a random forest algorithm also has some drawbacks.

- It requires much computational power as well as resources as it builds numerous trees to combine their outputs.
- It also requires much time for training as it combines a lot of decision trees to determine the class.
- Due to the ensemble of decision trees, it also suffers interpretability and fails to determine the significance of each variable.[7]

3. Decision Tree Regression Algorithm :

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems. Regression trees are needed when the response variable is numeric or continuous. Classification trees, as the name implies, are used to separate the dataset into classes belonging to the response variable. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.[8]

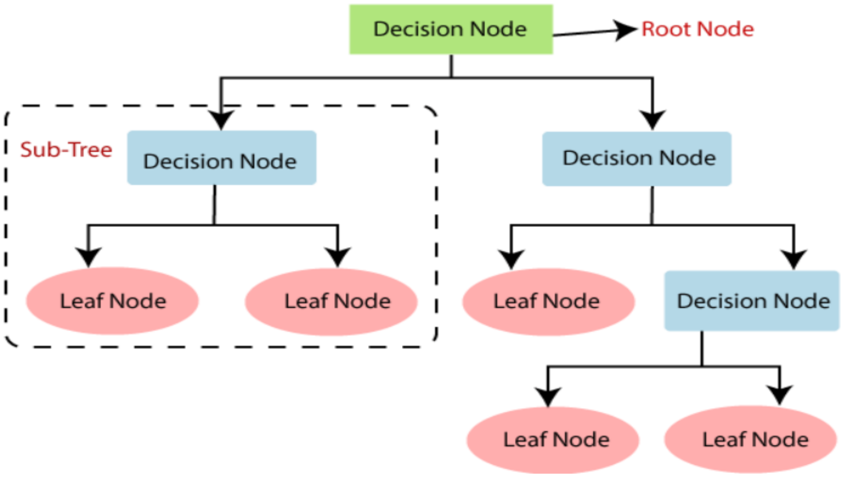
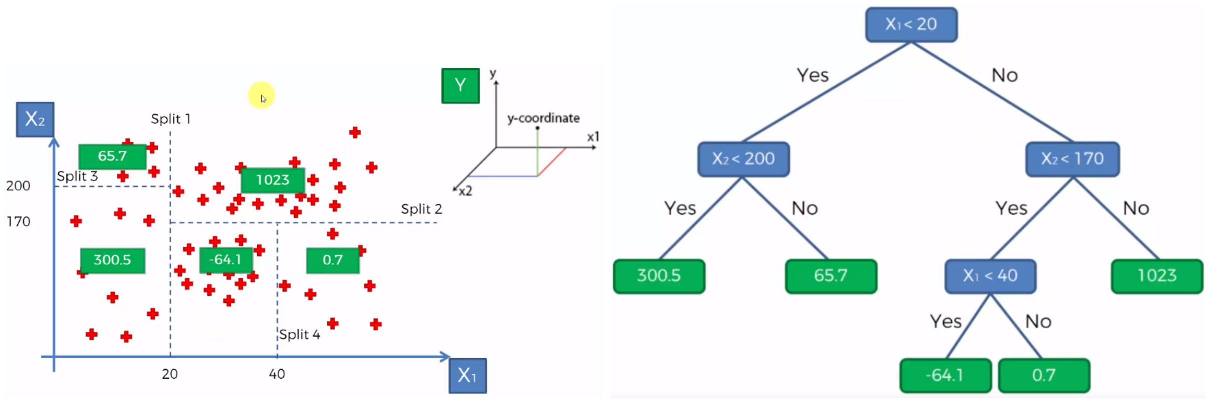


Fig 3 : Structure of a Decision Tree



(a) (b)

Fig 4 : Decision Tree Regressions

ASSESSMENTS

We will be making use of two Validation models, the Train-Test Split Validate and the K- Cross Validation , to assess and Validate our Selected Model.

1. Train - Test Split Validate:

This method is a validation technique used to assess the performance of the selected model. A given dataset is divided into three subsets i.e. , Testing dataset , Training dataset and the Validation dataset, each of these datasets have their different functions.

- Training Set : This dataset is used to train/ teach our machine learning model so that it can better understand how to apply technologies and give appropriate results.
- Validation Set : Is a small part of the dataset not used for training , so that it can give an unbiased evaluation of a model fitted , to the parameters.
- Testing Set : This is also a separate dataset from the training dataset that is used on the final fitted model , providing an unbiased evaluation of it and also how it applies in real time.

To split our dataset we'll be making use of Sklearn 'train_test_split' to split our dataset, and Fast_ml 'train_valid_test_split' to split it into the above mentioned three datase[9]t.

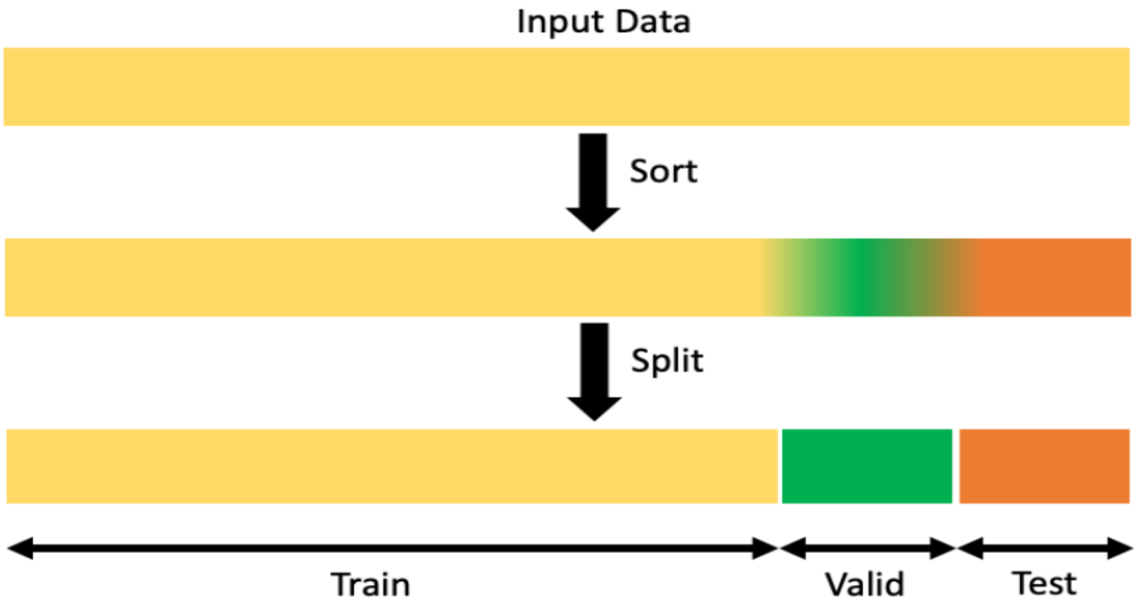


Fig 5 : Train_Valid_Test Split Validation

PRESENTATION AND VISUALIZATION

Our results will include :

1. Training Dataset
2. Testing Dataset
3. Visual Graphs and plots of the testing data , identifying the relations and dependencies of the different attributes of the dataset.
4. All the plots related to the machine learning algorithms being used.
5. Predicted Price attribute csv file.
6. Plots of any observed characteristics or trends , from the result.

ROLES

The team members are Soumya Ramesh and Shreya Roy. The work is project as data collection, that is finding the right dataset, data preprocessing, applying the machine learning models, applying the validation models and documentation. Dataset collection, data preprocessing will be done by Soumya Ramesh and validation models will be done by Shreya Roy. Application of the machine learning models and the document are done equally by both.

SCHEDULE

Date	Task
20/12/2021	LA Proposal
22/12/2021	Dataset Collection
24/12/2021	Data Preprocessing
06/01/2022	Prediction using Machine Learning Models
10/01/2022	Applying Validation Models
17/01/2022	Final Report Submission and Presentation

BIBLIOGRAPHY

[1]https://economictimes.indiatimes.com/industry/auto/cars-uvs/indian-used-car-market-to-grow-12-14-in-next-few-years/articleshow/87335065.cms

[2]https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9532845

[3]https://www.kaggle.com/avikasliwal/used-cars-price-prediction/activity

[4]https://www.javatpoint.com/linear-regression-in-machine-learning

[5]https://en.wikipedia.org/wiki/Random_forest

[6]https://levelup.gitconnected.com/random-forest-regression-209c0f354c84

[7]https://www.mygreatlearning.com/blog/random-forest-algorithm/#AdvantagesandDisadvantagesofRandomForest

[8]https://medium.com/pursuitnotes/decision-tree-regression-in-6-steps-with-python-1a1c5aa2ee

[9]https://towardsdatascience.com/how-to-split-data-into-three-sets-train-validation-and-test-and-why-e50d22d3e54c#:~:text=Train%2DValid%2DTest%20split%20is,of%20these%20datasets%20is%20below.