BIRMINGHAM CITY University

Car Price Predication

Using ML Algorithms

Subitted To: Besher Alhalabi

Submitted By: Shoaib Hassan

Roll No. 22193928

Contents

Abstract:
The world of Machine learning is unfolding the reality that it is energizing to obtain real-life use cases that are pertinent to our daily work. Artificial Intelligence and Machine Learning have convinced science that machine intelligence is now twirling around the world. Numerous applications of Machine Learning are virtual assistance, self-driving cars, health disorder detection and many others. Previous study shown that there are many AI applications already working in management field as risk handling in credit cards and in operational field as detection of prints and voices etc. This paper enlightens new way of prediction of data to help end user for the selection of best item in the market. For this purpose, an analytical AI algorithm has been introduced in this paper. This algorithm predicts some features of a used car by using data sets and calculates an accurate result for the end user. These results will help the buyer to make decision that which car is probably good in mileage and condition to buy.

Keywords: Machine Learning, Mileage, Artificial Intelligence, Algorithm, Dataset.

Introduction:
Machine learning and Artificial intelligence are the most trustworthy technologies in the present century. These techniques train our computer to act by itself and for taking the most accurate decisions without giving any command. These computers are highly accurate, efficient, and fast in producing results.

In the past few years, there has been a sudden increase in the production of cars. This non-neglectable amount of production is increasing the market value. There are some user-friendly online portals now available such as Cars and bids, Autotrader, CarsDirect and many others offering such a good path of communication between buyer and seller to be provided regarding trends and features that regulate the importance of a used car. Machine learning works on a certain set of features and predicts the wholesale value of a car.

Data Set Description:
The proposed algorithm considers the dataset consisting of the following features.

Name: Name of the specific car (identity).

Year: The time when the car been invented.

Selling price: Cost of the car that seller is demanding from the buyer.

Km-driven: The total distance in km's that car has travelled from the day of invention to the day of selling.

Fuel: Which fuel the car is using from Diesel, LPG, or Petrol.

Seller Type: Our data set includes three possibilities. If the status is Individual, this means there is only one person who is responsible to sell the car. This one person is the owner of the car. If the status is Dealer, it indicates that the owner(seller) is involving the third party for dealing with the buyer. When the status is Trustmark, then the seller is using some online portals to sell the car.

Transmission: The car is manual or automatic.

Owner: The first buyer of the car when the car is newly purchased is the first owner. Then if this first owner sells his car to the next person as a used car, that next person is the second owner and so on.

Mileage: Km distance covered by the car per liter.

Engine: Consumption of fuel as input and power per hour as output.

Maximum power: Maximum power of engine in horsepower.

Torque: The power that a new car generates at the first start.

Seats: How many seats are available in the car.

We will predict the price by using these attributes for Machine learning algorithm.

There are some other factors that need to be considered. For example, the oldest model car will have the less price of selling. This algorithm can be used in the perspective of comparison as well. It will compare the data and will show the result that which year model is best for the buyer.

Domain Description:

Before going toward implementation stage, first we need to discuss and overview about automobile domain. The Automobile or Automotive domain is the automobile industry. This industry designs, manufactures and sells vehicles.

The first petrol engine automobile with horse-drawn carriage fitted with engine is manufactured in 1885 by Karl Benz in Germany. Benz was permitted for patent in 1886 and then he started the first production of automobiles in 1888. After Benz, his wife had a long-distance trip in august 1888 to prove that this vehicle is suitable for daily use.

After this event, People started random experiments on vehicles and a vast number of models and types of vehicles started introducing in different markets. People were taking interest in the vehicle market that increased the market value proportion. This up ratio leads towards the manufacture of more advanced vehicles with multiple features available in a single motor car. There were 806 million cars in the world in 2007. The number of total cars in the world is increasing to one thousand million cars these days. India is one of the most highlighted countries for the car business in past two years. This sale touches the number of forty millions of car businesses till the present day. It is obvious that this rate will be increased in future due to the strong domestic market. This business engages the new generation's interest. Therefore, people are demanding more features in newly built cars. Many of these have VR technology that is very intellectual for the new generation.

Research on the automotive industry is just a few years older than a century. Major changes are expected in future in this industry because of customer demand, climate change and increment in fossil fuel prices due to lack of oil reserves. Some historians claim that in 1600, sail mounted carriage is invented as vehicle other than animal or human. It is also claimed by some historians that the key starting point of vehicle was invented in 1600 as "engine". After the innovation of the four-stroke internal combustion gasoline-fueled engine in 1876, the first

development of motor vehicles and first inauguration of vehicle firm established in Europe and America. The first petrol engine automobile with horse-drawn carriage fitted with engine is manufactured in 1885 by Karl Benz in Germany. Benz was permitted for patent in 1886 and then he started the first production of automobiles in 1888. After Benz, his wife had a long-distance trip in august 1888 to prove that this vehicle is suitable for daily use.

Development of other technologies such as steering wheels and floor mounted accelerators are introduced in the 1890's and in the early 1900's. This development in automotive made vehicles easy to use. In 1906, the famous vehicle model Ford's model T was developed. This era is also remembered for the better contemporary designs of vehicles. In the 1940's, during World War II, the automotive industry started building military vehicles. This new development leads the market towards building new strategies. 1950's to 1960's was the era of new looks of automobiles. It includes fiber glass bodies, vehicle comfort, emerging safety, speed limits, seat belts and heating and ventilation equipment. This era has been thunderous for the manufacture of light duty motor vehicles. The growth of industry revenue was exceptionally low as compared to past years. The global economic crisis that starts in 2007, accelerate financial troubles for many of the world's largest automotive industries. This industry is now again rising after 2009 due to high demand for advanced vehicles.

There are some manufacturing sectors acting as pillars for automotive domain. These manufacturing industries are working for advanced engines, brake systems, spare parts of cars, metal stamping, and others. This industry serves for retailing of new and used motor cars and components. This industry also deals with spare parts through dealers if car owner needs some components for his car.

There are four key business drivers that affect the automotive domain. Economic condition is the first key driver. When economic conditions are acceptable, people are more likely able to purchase new cars that gives momentum to the domain. The fall of the economy leads to the fall ratio of purchasing new vehicles. The automobile domain needs to balance the economy. Industries plan their capacities based on the prediction pf selling which is dependent on economic situation.

Consumer interest, preferences and demands are the second key drivers. There is an extreme awareness in advanced features. People are now demanding new varieties for sounds, shapes, looks, comfort, and fuel consumption storages. Vehicle safety is the most important for the customers these days and people are ready to pay for security additions.

Globalization is the third key driver. Global manufacturers are having global competitions that divide automobile makers into three layers. The first layer is GM, Toyota, and Honda. The second and third are attempted to consolidate or merge with other layers to compete with the first layer companies.
Technology is the fourth key driver. Every industry is working hard to seek innovative technologies to compete with other companies. The latest technology will offer the high price rates that impact directly on the market. These technologies include GPS, entertainment features, and satellite radio that are exactly according to customer demand.

Government is the fifth and last key driver. The government has applied strict regulations to deal with fuel economy in automobile manufactures. This driver set the standard that is compulsory for all the automobiles sold in different countries. This impacts greatly on global

vehicle manufacturers as they must update the products and price they are sold in different countries in the world.

Problem Definition:

The business of selling and purchasing used cars is now getting hype these days. Some people first do research, and they get information that which car is good to buy for them in their budget. But some people are not familiar with the features and quality of a car. These people then get the help of a car dealer. A car dealer is a person who serves his customer with different options of used cars. For dealing purpose, Dealer will charge extra payment and take commission from the purchasing as well. Customer might get confusion and make a wrong decision due to lack of information. This is our problem statement.

We are presenting a solution as Machine Learning Algorithm using supervised and unsupervised techniques that automatically shows the results of the used car according to the customer's need. This will save the time and energy of the customer. This will help a customer not to pay extra to the dealer. This algorithm will draw the most accurate results for the user. So, this solution presents the time saving, energy saving and money saving results. This platform will be very easy to use. Customers just need to add their desired year, model, color, required mileage, condition and some other requirements in specific columns. He will get the appropriate results in just a few seconds.

Data set exploration (preprocessing and wrangling):

We have a dataset of used cars. This dataset includes different columns with different features of a used car. We took features that are most considered by customers.

Dataset Exploration (preprocessing and wrangling).

Initially, we take raw data with missing values and errors. Before applying algorithms, we need to shape our data in a structure. For this purpose, we take some steps and strategies to make our data useful. This process is called Pre-Engineering or EDA (Exploratory Data Analysis). This process is very helpful to refine very huge datasets in an understandable form. This refined form will error less. This is the initial stage of algorithm creation.

In this procedure, the first step is to load our data set into a data frame. We are using python as a data frame. Before loading our data, we need to add requisite libraries in python. The next step is to build compatibility with our data. We completely understand the data and its features in this step, that is what it is and why we are using this kind of data in our algorithm creation. We will arrange our raw and random data on a table that will define every detail

	name	year	selling_price	km_driven	fuel	seller_type	transmission	owner	mileage	engine	max_power	torque	seats
0	Maruti Swift Dzire VDI	2014	450000	145500	Diesel	Individual	Manual	First Owner	23.4 kmpl	1248 CC	74 bhp	190Nm@ 2000rpm	5.000000
1	Skoda Rapid 1.5 TDI Ambition	2014	370000	120000	Diesel	Individual	Manual	Second Owner	21.14 kmpl	1498 CC	103.52 bhp	250Nm@ 1500-2500rpm	5.000000
2	Honda City 2017-2020 EXi	2006	158000	140000	Petrol	Individual	Manual	Third Owner	17.7 kmpl	1497 CC	78 bhp	12.7@ 2,700(kgm@ rpm)	5.000000
3	Hyundai i20 Sportz Diesel	2010	225000	127000	Diesel	Individual	Manual	First Owner	23.0 kmpl	1396 CC	90 bhp	22.4 kgm at 1750-2750rpm	5.000000
4	Maruti Swift VXI BSIII	2007	130000	120000	Petrol	Individual	Manual	First Owner	16.1 kmpl	1298 CC	88.2 bhp	11.5@ 4,500(kgm@ rpm)	5.000000

Figure 1: Data-Set Representation

separately. This table will also define the data set type although its string, float, categorical or integer etc.

The next and necessary step is to eliminate null and missing values. There are different methods to remove errors. As we are taking median of rows to remove errors in seats column in the figure given below.

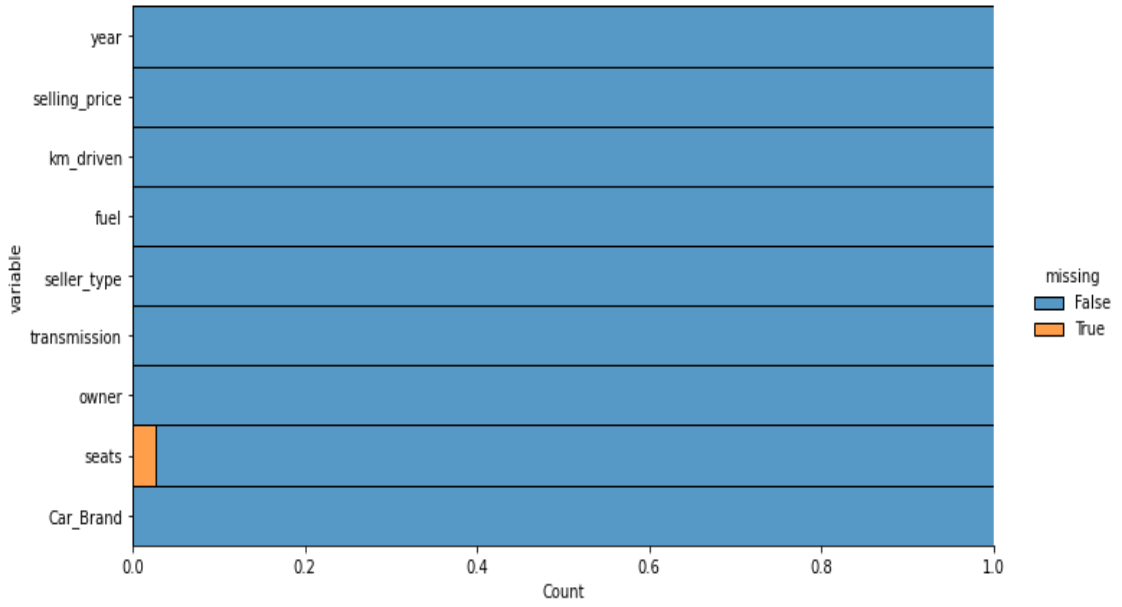

Figure 2: Show error values

For example, if the error is in row 5, we will take the median of row 4 and row 6. The result of median will be replaced with the error. We also apply transpose technique in our data during pre-processing when needed. Transpose means to exchange rows into columns and columns into rows. This technique is very helpful for efficient results.

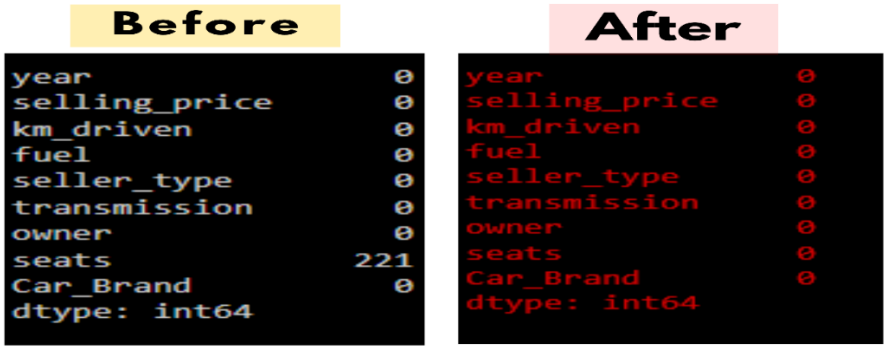

Figure 3: Removing error values

Sometimes, we exclude the whole row that contains the error. As we got errors in milage, torques, engine and maximum power. After studying data, we find out that these features are not very important for our algorithm and results as we have other important features in our

data so we just eliminate these features to get error free results. This elimination will not impact on our results. After this step, we will get a refined and understandable form of data.

We use a column named Car-name that contains some special characters with the name of car. These special characters could be the reason of errors. So, we just eliminated these special characters, and we only used the name of a car as Toyota, corolla etc.

We add a column of car-age to determine the total age of a car. This age can be calculated by the year of building and the present year. We add this column to have the most suitable results.

Now we visualize our data by creating bar graphs. This will be the more representable and understandable view of data. We will now study the given bar chat to compare some features to check the high priority customer demand. In the figure below, we compare three features with the selling price. We compare different fuels (Diesel, Petrol, CNG, LPG) with the selling price. According to the bar chat, we can clearly see that people are demanding diesel cars.

Then, the next comparison is between seller types with selling price. We have three selling types as Individual, Dealer seller type and Trustmark dealer. The results show that people trust dealer seller type more than the other two types.

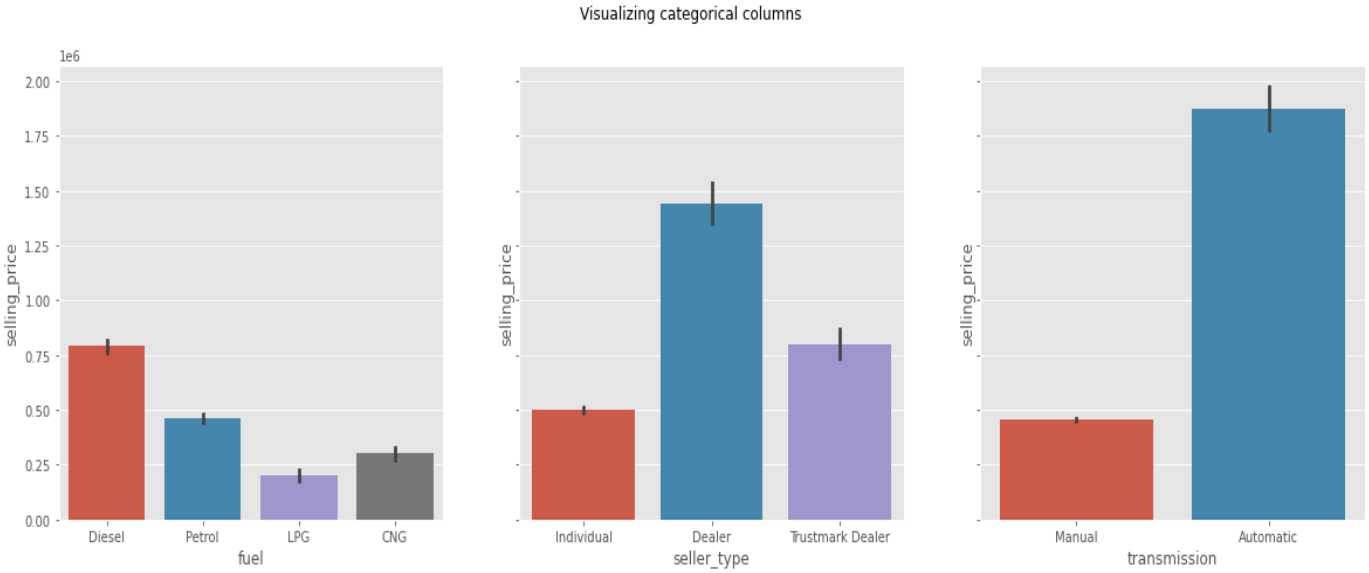

Figure 4: Show the individual values of features

The last comparison in this chart is between transmission and selling price. Transmission has two types as Manual and Automatic. Automatic cars are easy to drive and have more advanced features. Therefore, demand for automatic cars is higher than manual cars.

Our data is now refined and ready to process in an algorithm. Each algorithm has a scenario of its understanding. The algorithm we are introducing will understand and work only on integer values. So, we need to convert all the values (floating, categorical, etc.) of our table in integer form. For example, we have dealer type, car name and some other values other than integer values. We will assign some unique integer values to change the data type of this data

to make it understandable for the algorithm. If we do not implement this scenario, our algorithm might give bugs.

index	selling_price	km_driven	fuel	seller_type	transmission	owner	seats	Car_Brand	Car_age
0	450000	145500	1	0	0	0	5.0	Maruti	9
1	370000	120000	1	0	0	1	5.0	Skoda	9
2	158000	140000	0	0	0	2	5.0	Honda	17
3	225000	127000	1	0	0	0	5.0	Hyundai	13
4	130000	120000	0	0	0	0	5.0	Maruti	16

Figure 5: Change all datatypes into Integer data type

When we talk about data or datasets, there are different kinds and groups of datasets. Sometimes, we might have a huge amount of data. We need some techniques to categorize our data. The dataset used in our algorithm needs to be categorized because we have different kinds of it. We use Supervised ML algorithms and Unsupervised ML algorithms for our system. The supervised ML algorithm is used to process labeled data. Supervised ML is subcategory of Artificial Intelligence. It helps to organize the variety of real-world problems at some specific scale such as classifying spam folders at different places from your inbox. In this algorithm, we fed data as input into our system, and after validation process it generates an efficient and reliable output. This algorithm is used for categorizing and analyzing labeled datasets. For example, when we make graphs and charts. We define our data with names and values. So, this data is easy to recognize and properly defined. This kind of data is called labeled data. We classify our labeled data by using supervised ML algorithm. We have some real time examples for tagged or labeled datasets. We frequently see some sign boards on roads. Most of the signs are not understandable at first sight. The government labeled those signs with their names to make those signs understandable. So, this is the type of tagged data which can be categorized by using Supervisor ML algorithm. Examples of supervised ML algorithm are Linear Regression, lassco, Guassian NB, XGB classifiers which we are using in our system.

Supervised learning algorithms use data to train teaching models to produce the desired output. These training data sets are basically the inputs and accurate outputs that lead our system to run over time. This system measures accuracy until the error has been less.

Supervised learning works on two types of problems:

Data-Mining Classification: It uses an algorithm to correctly assign test data into specified categories. It acknowledges certain entities within the data sets used in the system and determines how these entities should be labeled or defined. Examples of classification algorithms are k-nearest neighbor, decision tree, and random forest.

Regression: This is used to understand the connection between dependent and independent variables. It is used to make projection purposes such as sale revenues for some specific business.

We use the following algorithms in our system:

Experiments:

Supervised Machine Learning Models:

Linear Regression Model:

This is the first and main model of our system of supervised learning. This model is used to predict our data and it produced the best percentage of accuracy as compared to other algorithms. This algorithm uses the values of a variable to find out the values of another variable. As we have two variables in linear regression model. One is dependent and the other one is independent variable. Dependent variable relies on input of the system. So dependent variable is output of the system expresses as 'y' and independent variable is an input of the system expresses as 'x'. We use labeled data in this model. If we talk about price predictions, this system provides efficient results. We use the linear regression expression to plot our data.

$$Y = a + Bx$$

Where X is the explanatory variable and Y is the dependent variable, B is the slope of the line and a is the intercept (the value of y when x = 0).

When we use this expression in our system, Y will be the selling price (output), x is the input (fuel, km driven, transmission).

This algorithm has two parts of processing. We use 70% of data for training and 30% of data for testing. Random state of data is 100. It will take 100 values of data randomly and then pick the data for training and testing according to the specified percentages.

The introduced model presents the prediction percentage as 82% that is very much efficient result as compared to other models.

It draws a hyper line between our inputs. Let suppose we have two inputs, A and B. This hyper line is drawn by using the general equation of straight line.

$$Y= mx + c$$

Where, m is the gradient of the line and c is the y-intercept (the point where the line crosses the y-axis).

The efficiency of the model depends on the projection of this line.

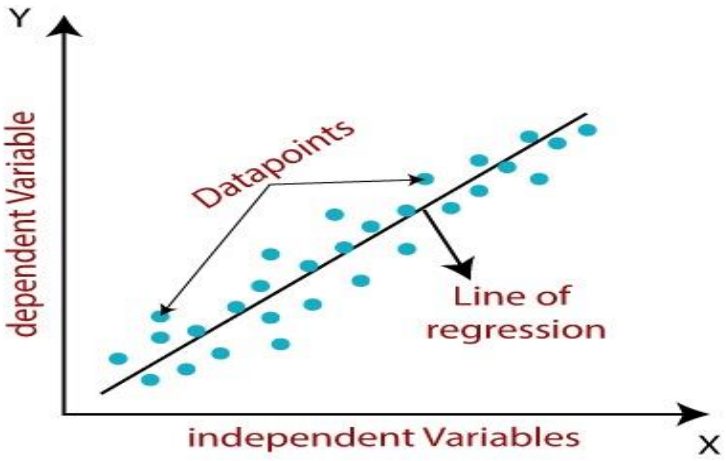

Figure 6: Linear Regression basic graph

Lasso Regression Model:

The word Lasso stands for Least Absolute Shrinkage and Selection Operator. Lasso Regression is the type of linear regression that uses shrinkage. Shrinkage is where the data values are shrunk to the mean (central point). This model is very simple as it uses very few parameters. This type of regression is useful when you want to automate the specific parts of model selection like parameter elimination or variable selection. The purpose of application of Lasso Regression is same as Linear Regression. This is also used for the prediction of data. This method uses random data for training and testing. We draw two lines for prediction, one for testing and the other for training. Training part is low biased that means the error occurrence rate is very less. The testing part got a higher error rate. Therefore, the performance of the training line is good. But the prediction percentage is 81% that is less than the linear regression prediction percentage.

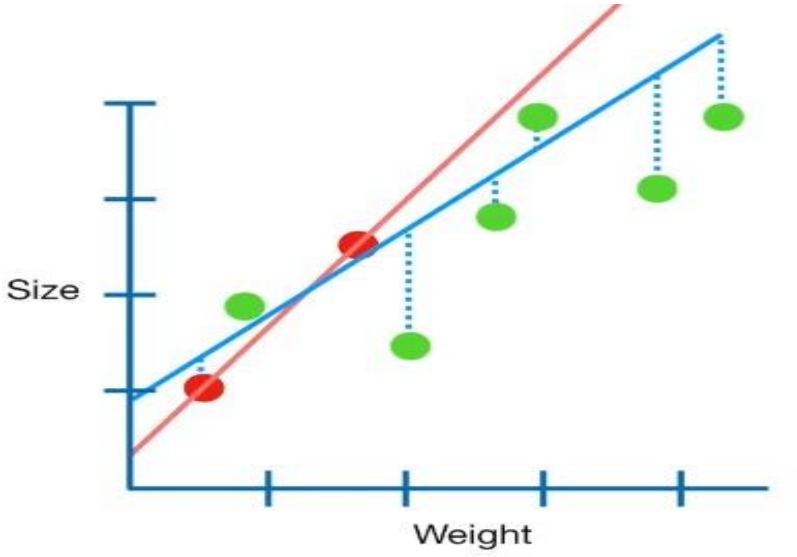

Figure 7: Lasso Regression Basic Graph

Bayes Theorem:

In statistics and probability theory, this theorem describes the probability of an event that is based on previous knowledge of features that might be related to an event. For example, if we consider the risk of arising health issues is increasing with age, then this theorem helps to find accurate results by applying condition on age and assuming that specific defective person is typical of the population as whole.

Following is the mathematical expression of the Bayes Theorem.

$$P(A|B) = \frac{P(B|A)P(A)}{\mathrm{P(B)}}$$

Gaussian Naive Bayes Model:

Gaussian Naive Bayes supports continuous features that contain values and models. It predicts some features by using already existing features.

This is the mathematical formula of Guassian Naive Bayas:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right)$$

According to the formula, we find a difference between our algorithm values and Guassian values. Our system values are on same point and have same values. Apart of this, Guassian values have a sudden rise and goes out of line during processing. This is why, this method is not appropriate for our system. We get 13.78% prediction percentage that could not be included in a reliable result.

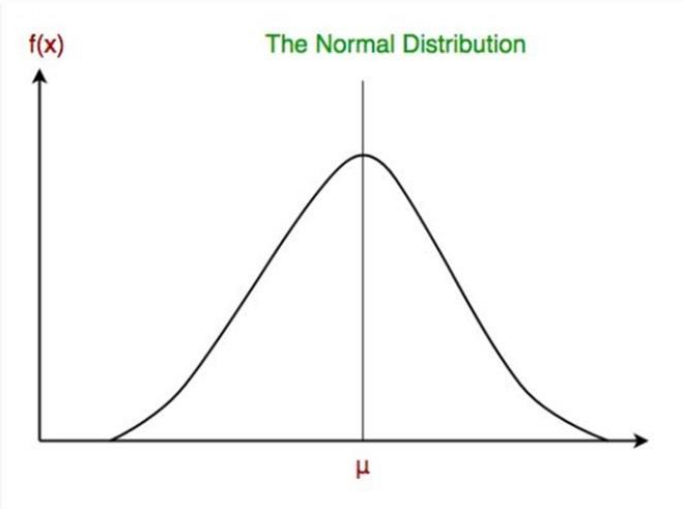

Figure 8: Gaussian NB basic graph

Un-Supervised Machine Learning Models:

The next algorithm is Unsupervised ML algorithm. This is exactly the opposite of supervised ML algorithm. This method is used to cluster or categorize unlabeled or untagged data. The kind of data that does not have any identification. This type of data set is mostly very difficult to understand. We do scale to measure the features of data to make this understandable. For example, if we see three different objects without knowing their names. Then we will measure their features to categorize those objects. Without identification datasets are called untagged datasets and the algorithm or technique that is used to classify unlabeled data is called unsupervised ML algorithm. So basically, this algorithm discovers patterns and then works on those patterns to draw results. In logical words, we give output to the system, and it gives us input.

Data Set Description:

	index	Maruti Swift Dzire VDI	2014	450000	145500	Diesel	Individual	Manual	First Owner	23.4 kmpl	1248 CC	74 bhp	190Nm@ 2000rpm	5
0	0	Skoda Rapid 1.5 TDI Ambition	2014	370000	120000	Diesel	Individual	Manual	Second Owner	21.14 kmpl	1498 CC	103.52 bhp	250Nm@ 1500-2500rpm	5.0
1	1	Honda City 2017-2020 EXi	2006	158000	140000	Petrol	Individual	Manual	Third Owner	17.7 kmpl	1497 CC	78 bhp	12.7@ 2,700(kgm@ rpm)	5.0
2	2	Hyundai i20 Sportz Diesel	2010	225000	127000	Diesel	Individual	Manual	First Owner	23.0 kmpl	1396 CC	90 bhp	22.4 kgm at 1750-2750rpm	5.0
3	3	Maruti Swift VXI BSIII	2007	130000	120000	Petrol	Individual	Manual	First Owner	16.1 kmpl	1298 CC	88.2 bhp	11.5@ 4,500(kgm@ rpm)	5.0
4	4	Hyundai Xcent 1.2 VTVT E Plus	2017	440000	45000	Petrol	Individual	Manual	First Owner	20.14 kmpl	1197 CC	81.86 bhp	113.75nm@ 4000rpm	5.0

Figure 9: Un-label data-set

K-Means Cluster:

Common unsupervised ML algorithms are k-means, Guassian mixture models and hierarchical. There are some applications of Unsupervised Machine Learning. We use unsupervised machine algorithms in visual perception e.g., object recognition in the computer field.It is used in image detection, classification, and segmentation by using radiology to detect diseases in the medical field.

Google uses unsupervised learning to classify articles on the same specific story from various news terminals. For example, the selection of president could be categorized under the label of “National News”.

As we have a huge dataset. This dataset was very random and irregular. The good method of applying algorithms and to get reliable and efficient results, we must need to arrange our datasets in a good, presented way.

Silhouette Coeficient Method:

The most common and popular unsupervised algorithm is K-Means clustering. It has some steps. In the first step, we define the number of clusters as ‘K’. Centroids of the clusters randomly choose K data points. We use “Silhouette Coefficient method” for finding the value of “K”.

Figure 10: Find K values Using silhouette coefficient

Next, classification of data based on Euclidean distance to either of the clusters. Then, by taking means of data point, we will update the centroids in each cluster. This algorithm seems simple but it has some keen steps to be taken. This includes choosing the points to initialize centroids in a correct way. This whole process is basically used for plotting points and recognizing the object. In our system, the object is our data set. Initially, we have random data without classification. We need an identification of this data. For this purpose, we used an unsupervised ML algorithm.

We process this untagged data in our system. For example, We will give data as input e.g. Price, fuel, seats, engine and year etc. Our algorithm then identifies the data in different specific columns. Our system separates our data into clusters by following K-mean steps. For example, we take the data Year and Price in the system. It will make two clusters and will add 2013, 2014, 2020 in cluster 'Year' and 2000$ and 3000$ in cluster 'Price'.

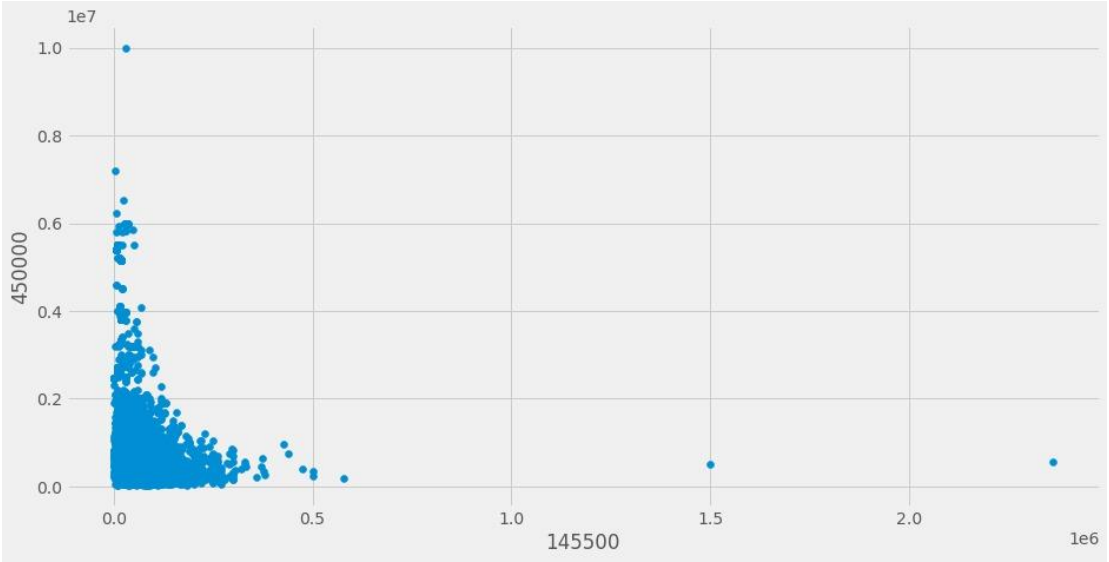

Figure 11: Cluster showing

We took 3 clusters for training and data deployment. Which is train our model according to the desired model and predication. It's work until unless to train and settle all the data in accordingly.

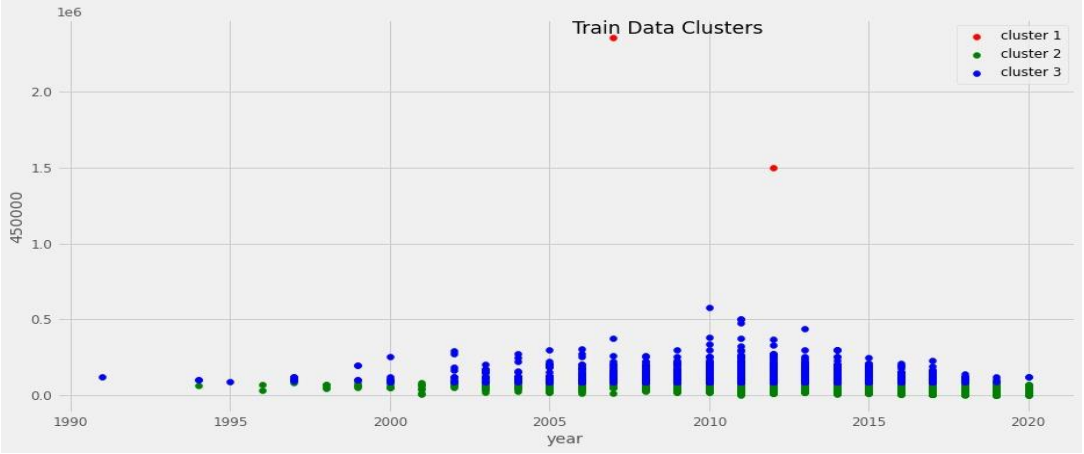

Figure 12: Training clusters

In classification problems, SVR (Support Vector regression) is a common algorithm in machine learning. SVR creates end dimensional vector space. It draws a hyper dimensional or hyper plain line at upper or lower position of the data. Then, it calculates the distance from line to the data. These hyper lines are called Kernels.

We have three types of Kernels.

1. Gaussian Kernel: Gaussian Kernels are universal Kernels. We do not have prior knowledge of the data in this Kernel. Which means we don't have the basics of data available.
2. Linear Kernel: A kernel that contains linear data is called Linear Kernel. This Kernel is used when the data is separable by using a single line. It is one of the most common Kernel to be used.
3. Polynomial Kernel: Polynomial Kernel represents the similarity of training samples that are called vectors in the feature space.

K-Means + SVR:

We use K-mean regressions + SVR is our system. When we do clustering by using K-mean regression and plot dots, at this point, SVR algorithm activates and draws hyper lines around these dots to create boundaries. It will make limitations and shrink the data to proceed data toward training. This mechanism is very useful to get more efficient results.

The following graph is the prediction graph for prices as output.

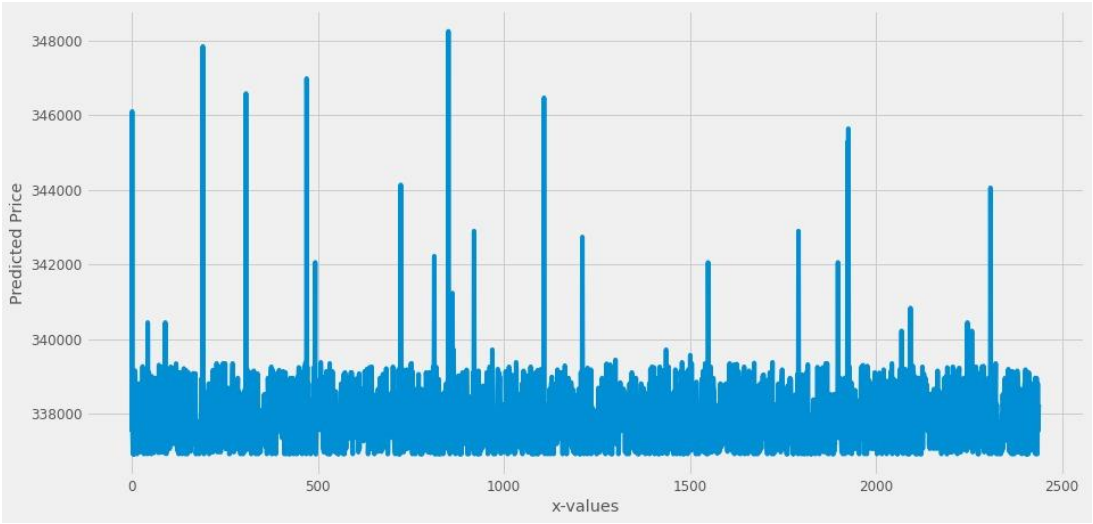

Figure 13: Predicated values

K-means clustering is mostly used when we do not have specific desired outcome variable that we are trying to predict. Instead, it is also used when we have a couple of features, we use to find collections of observations that share similar characteristics.

Analysis and Results:

Supervised ML Algorithms Analysis

S.NO	**Linear Regression**	**Lasso Linear Regression**	**Gaussian Naïve Bayes**
1	**82 %**	**81%**	**13.78 %**

Un-Supervised ML Algorithm Analysis:

S.NO	**MAE (Mean Average Error)**	**MSE (Mean Square Error)**
1	412264.32	858394.76

Summary:

This is the era where people are moving towards advancement but in the easiest and most efficient ways. In the past few years, we have seen a great rise in retail of used cars. If someone wants to upgrade his car model, he for sure can buy a new model but at the same time, he can buy the same model in used condition. This used car may be in very good condition, just like the new one, and you need to pay a very reasonable amount as it is difficult for some people to pay for new cars. To find used cars in good conditions in the market is a time taking process. People trust more in used cars instead of buying new ones. Buying used cars can save a person's money and time. People can buy old cars by visiting directly to the used car show rooms, by contacting an agent and buy using online applications. The system we proposed is going to offer you one clicks information about used cars. This is basically an algorithm that will ask you about your requirements and will provide the best results according to your demand. This algorithm is called Machine Learning Algorithm and uses some supervised and unsupervised techniques to arrange, analyze and process the data sets. We use some common features of cars in a dataset. A person just needs to enter the specific data, for example the required engine, mileage, color of the car, seats, fuel, and price range. The system will take this data as input and after processing, it will serve you with the best results as output. This system is very helpful if someone has no awareness and knowledge of buying cars. This system is very useful for the person who has addiction of changing models and colors in cars.

Conclusions:

In this report, we describe a Machine Learning algorithm with some methods and techniques that is introduced to serve people the easiest ways of buying used cars. The description is based on the emerging demand for used cars. We hope this idea will help interested buyers to get more knowledge about cars and features. This work can be further studied for adding more advanced features and for the improvement of the prediction percentage.

References:

1. Ullah, R., Hayat, H., Siddiqui, A.A., Siddiqui, U.A., Khan, J., Ullah, F., Hassan, S., Hasan, L., Albattah, W., Islam, M. and Karami, G.M., 2022. A Real-Time Framework for Human Face Detection and Recognition in CCTV Images. Mathematical Problems in Engineering, 2022. pp. 1–12. Available at: https://doi.org/10.1155/2022/3276704.

2. Mou, A.D., Saha, P.K., Nisher, S.A. and Saha, A., 2021, February. A comprehensive study of machine learning algorithms for predicting car purchase based on customers demands. In 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD) (pp. 180-184). IEEE. Available at: https://doi.org/10.1109/icict4sd50815.2021.9396868.

3. Jayakameswaraiah, M. and Ramakrishna, S., 2014. Design and Development of Data Mining System to Analyze Cars using Improved ID3 with TkNN Clustering Algorithm. (IJCST), 2(4).

4. Durgesh, K.S. and Lekha, B., 2010. Data classification using support vector machine. Journal of theoretical and applied information technology, 12(1), pp.1-7.

5. Pudaruth, S., 2014. Predicting the price of used cars using machine learning techniques. Int. J. Inf. Comput. Technol, 4(7), pp.753-764.

6. Venkatasubbu, P. and Ganesh, M., 2019. Used Cars Price Prediction using Supervised Learning Techniques. Int. J. Eng. Adv. Technol.(IJEAT), 9(1S3).

7. Mahesh, B., 2020. Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet], 9, pp.381-386.

8. Feitosa, M.F., Rocha, S., Gonçalves, G.D., Ferreira, C.H. and Almeida, J.M., 2022, November. Sentiment Analysis on Twitter Repercussion of Police Operations. In Proceedings of the Brazilian Symposium on Multimedia and the Web (pp. 84-88). Available at: https://doi.org/10.1145/3539637.3558050.

9. Delua, J. (2021). Supervised vs. unsupervised learning: What's the difference? [online] IBM. Available at: https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning.

10. Carzato (2022). Demand for Used Cars Has Dropped Compared to Last Year, But Prices Remain High. [online] Carzato. Available at: https://www.carzato.com/demand-for-used-cars-has-dropped-compared-to-last-year-but-prices-remain-high/ [Accessed 10 Jan. 2023].

11. Pal, N., Arora, P., Kohli, P., Sundararaman, D. and Palakurthy, S.S., 2018, April. How much is my car worth? A methodology for predicting used cars' prices using random forest. In Future of Information and Communication Conference (pp. 413-422). Springer, Cham.

Appendixes:

Main Code:

I have the single Co-lab Notebook file of code in which the both Supervised and UN-Supervised ML algorithms contain. Click-Here for the Co-lab Notebook file.

Data Set:

We have two datasets. Please download both and upload.

Labeled data-set for Supervised ML algorithms. Click-Here for download.

Un-Labeled data-set for Un-Supervised ML Algorithms. Click-Here for download.