Used Car Price Prediction using Machine Learning

Jimmy Aghera
Dept. of Computer Science and Engineering
Depstar- Charusat
Anand, India
jimmyaghera123@gmail.com

Rudra Barad
Dept. of Computer Science and Engineering
Depstar- Charusat
Anand, India
rudra.barad@gmail.com

Abstract - This research paper presents car price prediction system by using the supervised machine learning technique. This research uses different algorithms of Machine Learning which offers highest accuracy of 90%. We applied three machine learning algorithms which are Linear Regression, Random Forest, and Naïve Bayes out of which highest accuracy was obtained in Random Forest. We also attempted to create a mathematical model that could predict the price of a used car based on previous consumer data and a collection of characteristics. The data used for the prediction was collected from the Kaggle.

General Term
Machine Learning

Keywords
Car Price Predictions, Linear Regression, Random Forest, Naïve Bayes

I. INTRODUCTION

In the last decade, car production has gradually increased, with over 23 million cars expected to be manufactured and sold by 2020. As a result, the used car market has developed into a thriving industry in its own right. The new coming of online entrances has worked with the requirement for both the client and the vendor to be better educated about the patterns and examples that decide the estimation of a trade-in vehicle on the lookout. The rise of online entryways like CarDheko, Carstrade, Carwale, Cars24, and numerous others has worked with the requirement for both the client and the dealer to be better educated about the patterns and examples that decide the estimation of the pre-owned vehicle. Machine Learning algorithms can be used to estimate a car's retail value, based on a certain set of features. There is no particular unified algorithm for calculating the used car selling price, as different websites use different algorithms to produce it. Without directly entering the data into the desired website, one can quickly get an approximate approximation of the price by training mathematical models for forecasting prices. The main objective/goal of this paper are:

1. Design - The study involves the development of a method that explains the linear association between the variables X and Y, which are price and other variables such as car model and make.

2. Predict: The study forecasts the vehicle's price using a linear regression model that recognizes and projects various trends and estimates the car's value.

3. Confirm: The study determines which vehicle-related variable is the most accurate predictor of its price.

The arrangement of this paper is such that Section II is Literature Review, Section III consists of methodology and analysis, and the findings are presented in Section IV. Section V brings the paper to a close by outlining potential study directions and including a list of references.

II. LITERATURE REVIEW

The study's main goal is to create a strong regression model that can accurately forecast car prices. Data mining is the method of collecting valuable data from a variety of sources. So, we used various algorithms for this purpose. Linear regression, naïve bayes, and random forest are algorithm which is used in this research and we had a small discussion and comparison between all the algorithm we used. We used Kaggle's dataset for our research.

Metadata of dataset used
Total columns: 12
(Name, Location,
Year, Kilometers_Driven, Fuel_Type, Transmission,
Owner_Type, Mileage, EngineSeats, Price, company)
Total rows: 5965
Total Items: 71,580
Null values: 53

III. METHODOLOGY AND ANALYSIS

To do so, we'll need some past data on used vehicles, for which we'll use price and other standard attributes. And then first job is to clean data, as some car may be irrelevant and can lead to wrong price prediction. And also we have to remove null value from the data for crystal clear output.

After cleaning data, we can use them for prediction but we have many algorithms for prediction we have used three of them namely (I) Linear regression (II) Random forest and (III) Naïve bayes. They are as follows:

(I) Linear regression

Regression is a technique for predicting a goal value using independent predictors. This method is primarily used for forecasting and determining cause and effect relationships among variables. The number of independent variables and the form of relationship between the independent and dependent variables are the main differences in regression techniques. Simple linear regression is a type of regression analysis in which there are only one independent variable and the independent(y) and dependent(x) variables have a linear relationship. The linear equation shown below can be used to model the line for linear regression is

$$y = x * a1 + a0$$

Now, the Cost function assists us in determining the best possible values for a0 and a1 in order to obtain the best possible fit line for the data points. For getting the best values of a0 and a1, we transform this search problem into a minimization problem in which we want to reduce the difference between the expected and actual values.

$$J = 1/n \sum_{k=1}^{n} (pred - y)^2$$

Gradient descent is the next crucial term to grasp in order to comprehend linear regression. Gradient Descent is a method of updating a0 and a1 to reduce the cost function. The idea is that we start with some a0 and a1 values and then reduce the cost by changing them iteratively. Gradient descent aids in the transformation of values.

(II) Random forest

The decision tree is used in the random forest, which is made up of a large number of individual decision trees that work together as an ensemble. Each tree in the random forest produces a class prediction, and the class with the most votes become the prediction of our model. Any of the individual constituent models will outperform a large number of relatively uncorrelated models (trees) working as a committee.

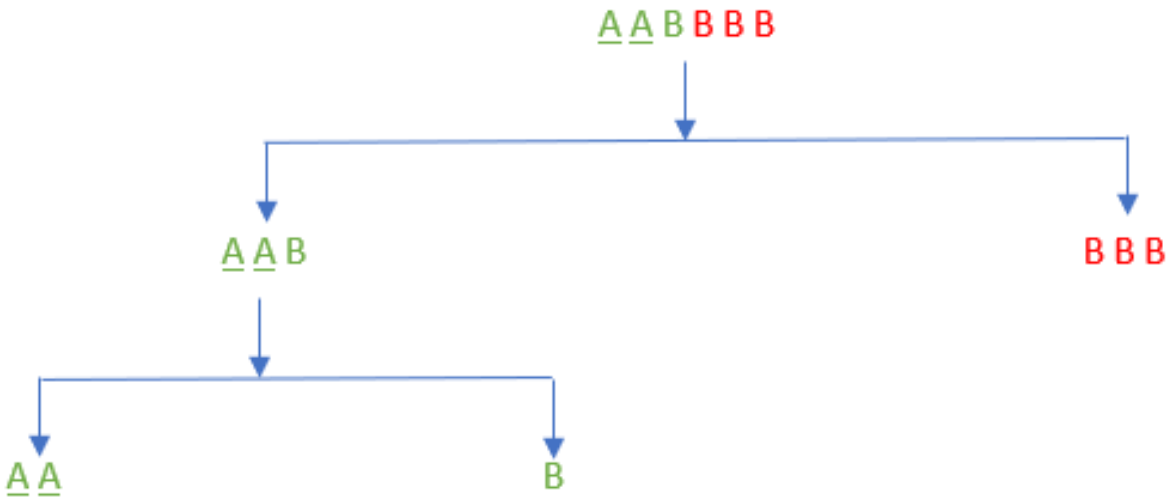

Figure 1 A single tree working in a random forest algorithm.

The secret is the low correlation between models. Uncorrelated models can generate ensemble predictions that are more reliable than any of the individual predictions, similar to how low-correlation investments (like stocks and

bonds) come together to create a portfolio that is greater than the sum of its parts. The trees shield each other from their individual mistakes, which results in this wonderful effect.

(III) Naive bayes

The easiest algorithm you can use to analyse your data is Naive Bayes. As the name implies, this algorithm makes an assumption that all of the variables in the dataset are "Naive," that is, uncorrelated. Naive Bayes is a widely used classification algorithm for determining the dataset's base accuracy.

$$P(c \mid x) = P(x \mid c)\frac{P(c)}{P(x)}$$

$P(c \mid x)$: posterior probability
$P(c)$: probability of class.
$P(x \mid c)$: likelihood which is the probability of predictor given class.
$P(x)$: prior probability of predictor.

So, by making graph we can see the irrelevant data. Table 1 below summarizes the data used in the system after cleaning which was taken from Kaggle.

	Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Seats	Price	company
0	Maruti Wagon R	Mumbai	2010	72000	CNG	Manual	First	26.60	998	5	1.75	Maruti
1	Hyundai Creta 1.6	Pune	2015	41000	Diesel	Manual	First	19.67	1582	5	12.50	Hyundai
2	Honda Jazz V	Chennai	2011	46000	Petrol	Manual	First	18.20	1199	5	4.50	Honda
3	Maruti Ertiga VDI	Chennai	2012	87000	Diesel	Manual	First	20.77	1248	7	6.00	Maruti
4	Audi A4 New	Coimbatore	2013	40670	Diesel	Automatic	Second	15.20	1968	5	17.74	Audi
...	...	...	...	...	...	...	...	...	...	...	...	...
5960	Maruti Swift VDI	Delhi	2014	27365	Diesel	Manual	First	28.40	1248	5	4.75	Maruti
5961	Hyundai Xcent 1.1	Jaipur	2015	100000	Diesel	Manual	First	24.40	1120	5	4.00	Hyundai
5962	Mahindra Xylo D4	Jaipur	2012	55000	Diesel	Manual	Second	14.00	2498	8	2.90	Mahindra
5963	Maruti Wagon R	Kolkata	2013	46000	Petrol	Manual	First	18.90	998	5	2.65	Maruti
5964	Chevrolet Beat Diesel	Hyderabad	2011	47000	Diesel	Manual	First	25.44	936	5	2.50	Chevrolet

5965 rows × 12 columns

Table 1 Shows the clean data of dataset

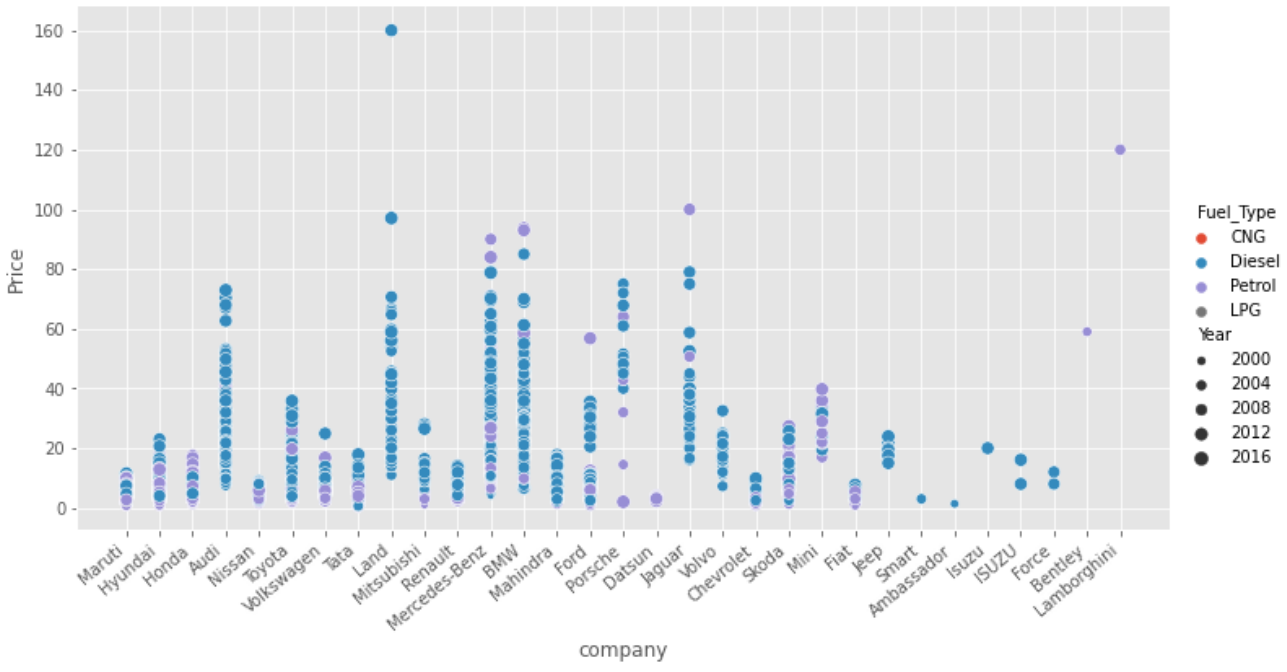

Figure 2 Graphical representation of Price to Company with Fuel type and year of manufacture

The *Figure 2* above show many things at time, firstly we can directly see that the Diesel car and patrol care are more than LPG and CNG we can see that in *Figure 6*, the size of the dot shows that how old the car is the clear representation of that is also shown in *Figure 7* from which we can conclude that as car are elder the price of car is decreased accordingly. And by analysing this we can see that is there any single car which don't follow this trend or not and we can easily find that out using graph. And we can remove that as that will deviate our algorithm and make some confusion in it.

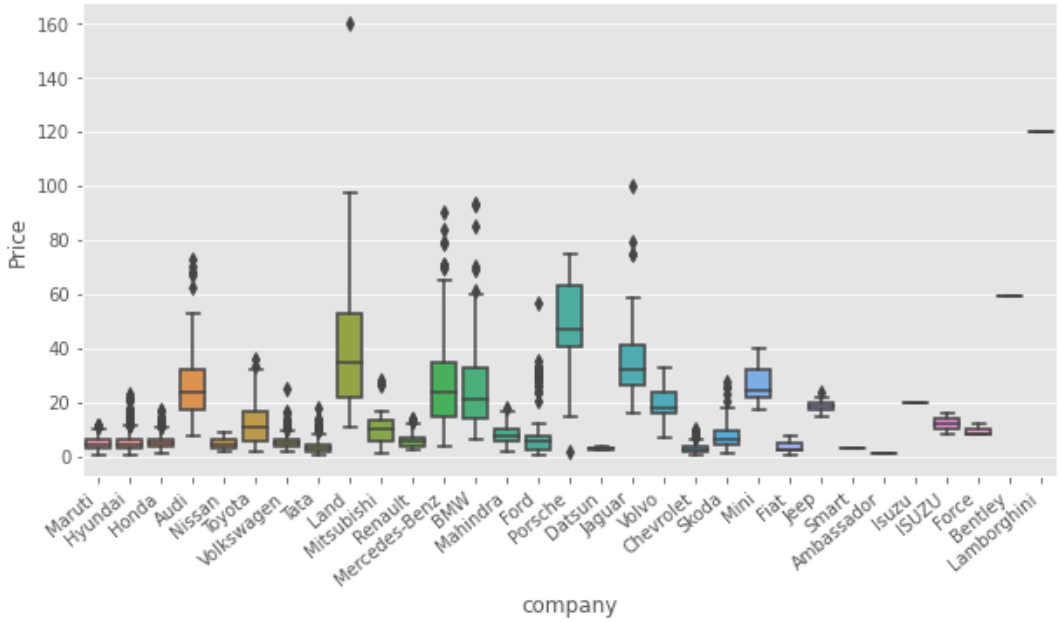

Figure 3 Box plot for seeing the density for Price to Company

From the *Figure 3* we can see that what are the density of cars when we plot graph of price verses company and we can conclude that which company car price is higher and then accordingly algorithm can understand that which company stands under expensive one or cheap one. And same way for engine verses company graph which is shown *Figure 4* in is also give in which we can see that the higher the engines CC more price of the car will be & more the expensive car company will be there.

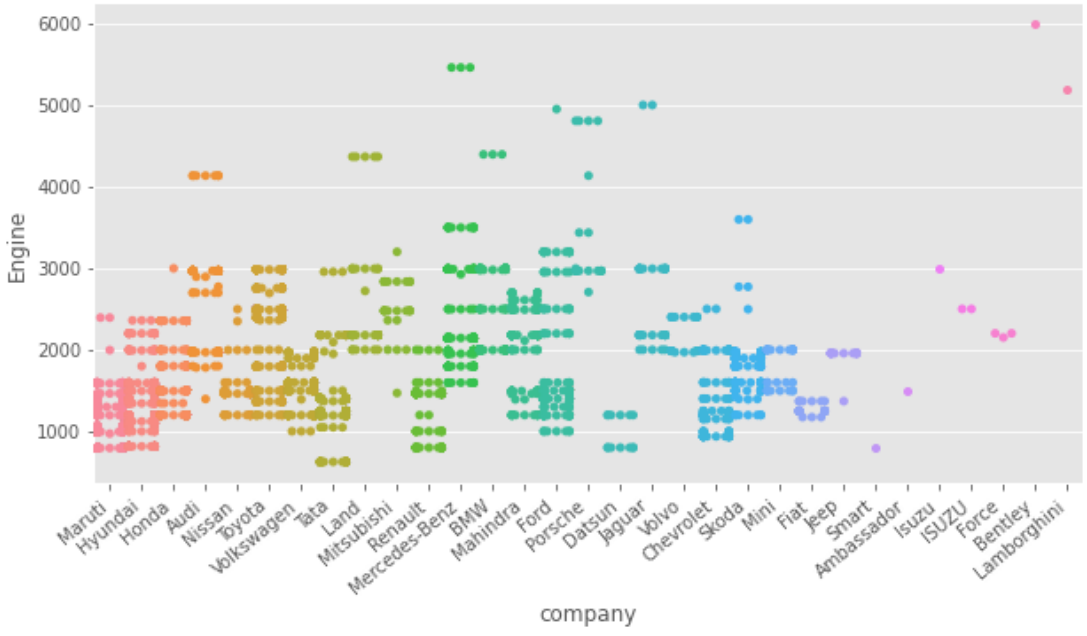

Figure 4 Engine power to company representation.

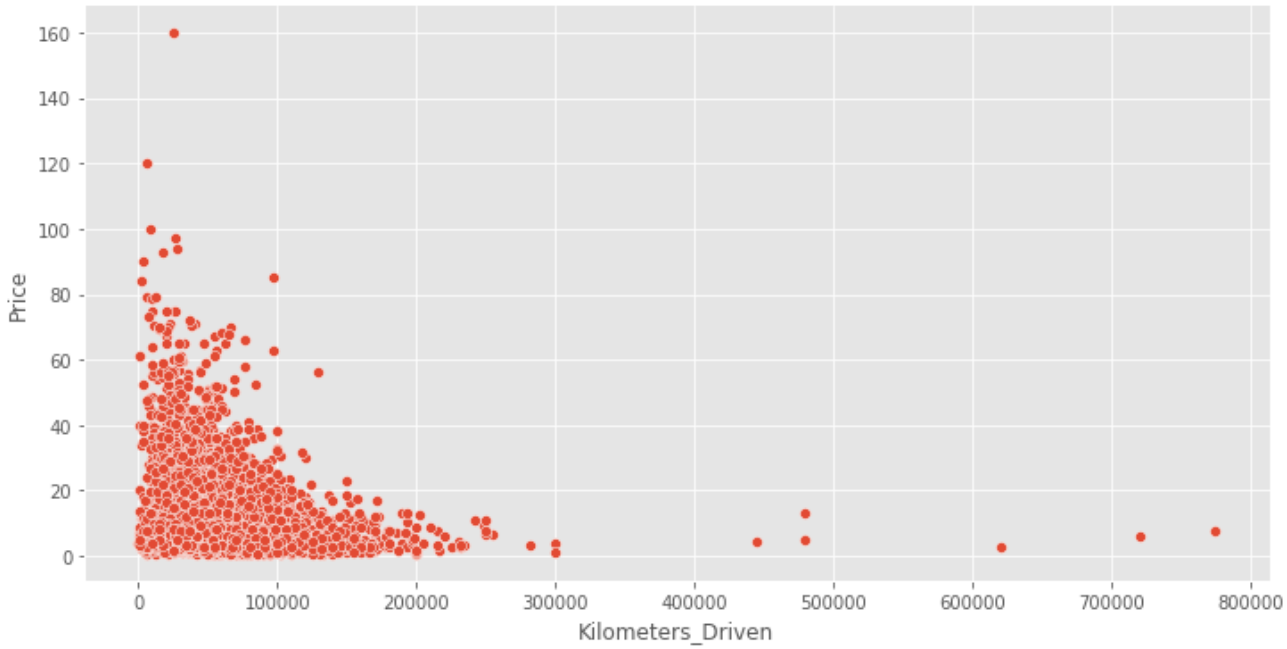

Figure 5 Price to Kilometres driven relation graph plot.

We can see the dot plot of price to Kilometres driven of the car, and by seeing the graph we can remove the unusual data which can mislead the algorithm. So, that better output can be provided to user. And also, algorithm can see that the car which is less driven have higher price compared to car having higher driven.

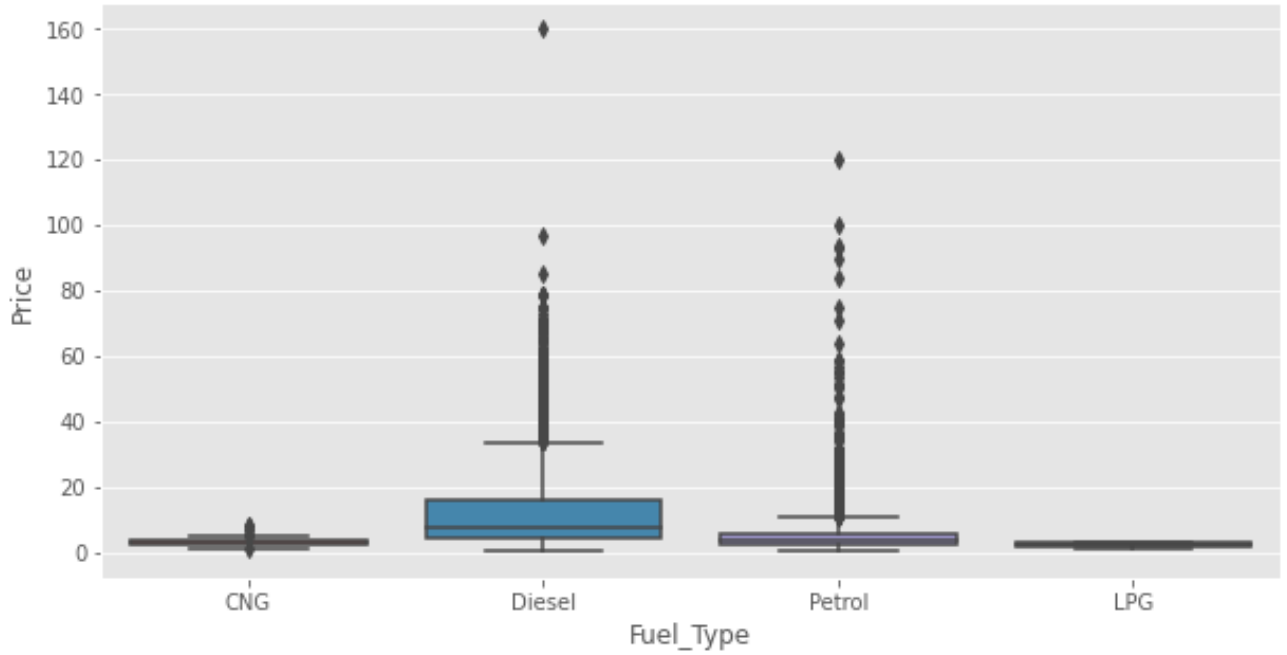

Figure 6 Price to Fuel Type box plot.

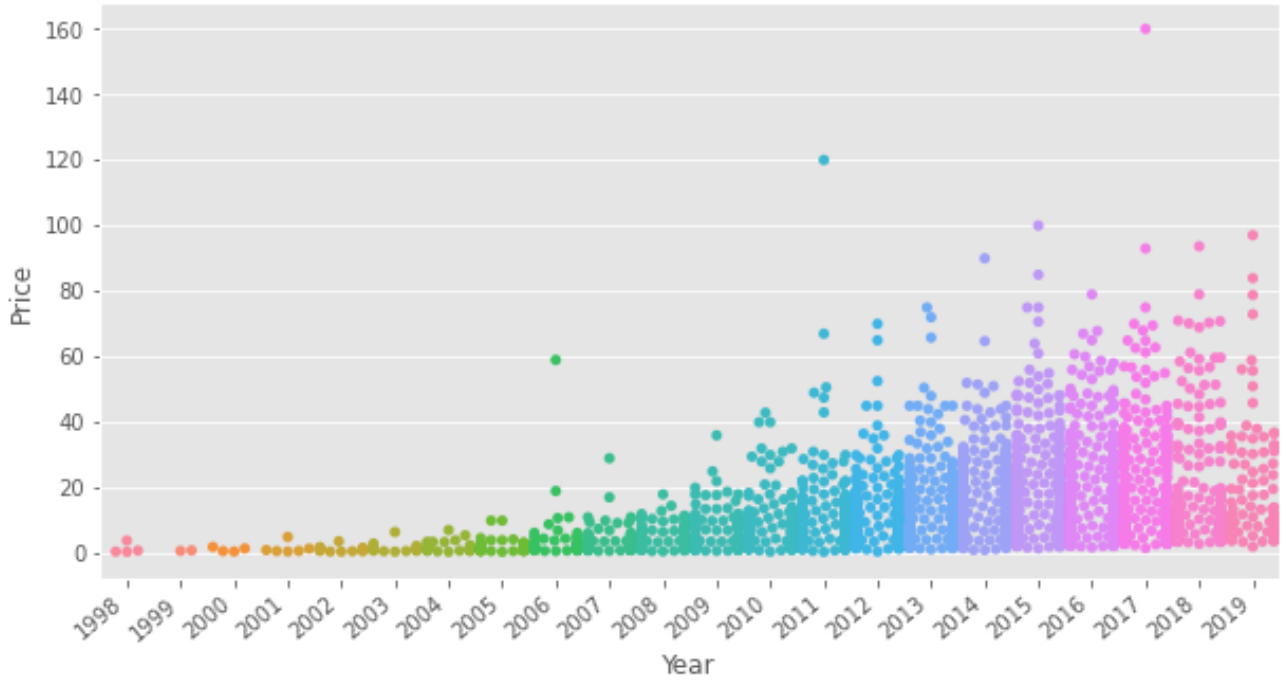

Figure 7 Price to manufacture year swarm plot.

IV. RESULTS

The least square method was used to estimate the model, and the following Minitab results were obtained.

R^2	Accuracy
Linear Regression	83.67%
Naïve Bayes	88.68%
Random Forest	89.01%

Table 2 R^2 Values

The percentage answer variance in a variable called R-square is explained by a linear model (Rsq). This means that a high R-square value indicates that the model is more suited to the data and hence produces more reliable results.

V. CONCLUSION

Due to the large number of characteristics that must be taken into account for an accurate forecast, car price prediction can be a difficult task. For conducting similar research using

different prediction techniques, the data set used in this research can be very proved very valuable. During the processing and analysis step, the proposed method evaluated variables and selected the most important put of the dataset, reducing the model's complexity by removing irrelevant variables. The collection and preprocessing of data is a crucial step in the prediction process. Data cleaning is one of the processes that improves prediction accuracy, but it is inadequate for complex data sets like the one in this study. Applying single machine learning algorithm on the data may not be fruitful therefore, we tried with multiple machine learning algorithms to gain accuracy sufficient enough to predict the price of car about 90%. Car prices may also be forecasted using the same or a different forecasting programmed. The data collected for this study aided in the estimation of used car prices using three separate machine learning algorithms, including Linear Regression, Naïve Bayes, and Random Forest. However, the proposed system has the disadvantage of consuming significantly more computing resources than a single machine learning algorithm. Despite this, this method has obtained outstanding results in the problem of car price prediction. The aim of future research is to see if this system will operate with a variety of data sets.

VI. REFERENCES

[1] Pudaruth,S. 2014. “Predicting the Price of Used Cars Using Machine Learning Techniques”, International Journal of information & Computation Technology,4(7), p.753-764.

[2] Noor, K., & Jan, S. (2017). Vehicle Price Prediction System using Machine Learning Techniques. International Journal of Computer Applications, 167(9), 27-31.

[3] 3.2.4.3.1. sklearn.ensemble.RandomForestClassifier — scikit-learn 0.19.2 documentation. (n.d.). Retrieved from: http://scikitlearn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html [accessed: August 30, 2018].

[4] Used cars database. (n.d.) Retrieved from: https://www.kaggle.com/orgesleka/used-carsdatabase. [accessed: June 04, 2018].

[5] Soejima Y, Hirose H, “Auction Price Estimation for Used Cars by Regression Methods (Competition 1)” in Proceedings of the Japan Society for Computer Science and Statistics (2011), pp. 9-12.

[6] Monburinon, Nitis, Prajak Chertchom, Thongchai Kaewkiriya, Suwat Rungpheung, Sabir Buya, and Pitchayakit Boonpou. "Prediction of prices for used car by using regression models." In 2018 5th International Conference on Business and Industrial Research (ICBIR), pp. 115-119. IEEE, 2018

[7] Gegic, Enis, Becir Isakovic, Dino Keco, Zerina Masetic, and Jasmin Kevric. "Car price prediction using machine learning techniques." TEM Journal 8, no. 1 (2019): 113.

[8] https://towardsdatascience.com/all-about-naive-bayes-8e13cef044cf

[9] https://towardsdatascience.com/understanding-random-forest-58381e0602d2

https://towardsdatascience.com/introduction-to-machine-learning-algorithms-linear-regression-14c4e325882a