Used Cars Price Prediction using Supervised Learning Techniques

Article *in* International Journal of Engineering and Advanced Technology · December 2019

DOI: 10.35940/ijeat.A1042.1291S319

Mukkesh Ganesh
VIT University

Used Cars Price Prediction using Supervised Learning Techniques

Pattabiraman Venkatasubbu, Mukkesh Ganesh

Abstract: The production of cars has been steadily increasing in the past decade, with over 70 million passenger cars being produced in the year 2016. This has given rise to the used car market, which on its own has become a booming industry. The recent advent of online portals has facilitated the need for both the customer and the seller to be better informed about the trends and patterns that determine the value of a used car in the market. Using Machine Learning Algorithms such as Lasso Regression, Multiple Regression and Regression trees, we will try to develop a statistical model which will be able to predict the price of a used car, based on previous consumer data and a given set of features. We will also be comparing the prediction accuracy of these models to determine the optimal one.

Keywords: ANOVA, Lasso Regression, Regression Tree, Tukey's Test

I. INTRODUCTION

The used car market is an ever-rising industry, which has almost doubled its market value in the last few years. The emergence of online portals such as CarDheko, Quikr, Carwale, Cars24, and many others has facilitated the need for both the customer and the seller to be better informed about the trends and patterns that determine the value of the used car in the market. Machine Learning algorithms can be used to predict the retail value of a car, based on a certain set of features.

Different websites have different algorithms to generate the retail price of the used cars, and hence there isn't a unified algorithm for determining the price. By training statistical models for predicting the prices, one can easily get a rough estimate of the price without actually entering the details into the desired website. The main objective of this paper is to use three different prediction models to predict the retail price of a used car and compare their levels of accuracy.

The data set used for the prediction models was created by Shonda Kuiper[1]. The data was collected from the 2005 Central Edition of the Kelly Blue Book and has 804 records of 2005 GM cars, whose retail prices have been calculated. The data set primarily comprises of categorical attributes along with two quantitative attributes.

The following are the variables used:

Revised Manuscript Received on December 02, 2019

***** Correspondence Author

Pattabiraman Venkatasubbu*, School of Computing Science and Engineering, Vellore Institute of Technology, Chennai, India. Email: pattabiraman.v@vit.ac.in

Mukkesh Ganesh, School of Computing Science and Engineering, Vellore Institute of Technology, Chennai, India. Email: g.mukkesh2017@vitstudent.ac.in

Price: The calculated retail price of GM cars. The cars which were selected for this data set were all less than a year old and were considered to be in good condition.
Mileage: The total number of miles the car has been driven
Make: The manufacturer of the car
Model: The specific models for each car
Trim: The type of car model
Type: The car's body type
Cylinder: The number of cylinders present in the engine
Liter: The fuel capacity of the engine
Doors: The number of doors in the car
cruise: A categorical variable (binary), which represents whether cruise control is present in the car (coded 1 if present)
sound: A categorical variable (binary), that represents whether upgraded speakers are present in the car (coded 1 if present)
Leather: A categorical variable (binary), that represents whether the car has leather interiors (coded 1 if present)

Using these attributes, we will try to predict the price by using the Statistical Analysis System (SAS) for exploratory data analysis.

II. LITERATURE SURVEY

Overfitting and underfitting come into picture when we create our statistical models. The models might be too biased to the training data and might not perform well on the test data set. This is called overfitting. Likewise, the models might not take into consideration all the variance present in the population and perform poorly on a test data set. This is called underfitting. A perfect balance needs to be achieved between these two, which leads to the concept of Bias-Variance tradeoff. Pierre Geurts [2] has introduced and explained how bias-variance tradeoff is achieved in both regression and classification. The selection of variables/attribute plays a vital role in influencing both the bias and variance of the statistical model. Robert Tibshirani [3] proposed a new method called Lasso, which minimizes the residual sum of squares. This returns a subset of attributes which need to be included in multiple regression to get the minimal error rate. Similarly, decision trees suffer from overfitting if they are not pruned/shrunk. Trevor Hastie and Daryl Pregibon [4] have explained the concept of pruning in their research paper. Moreover, hypothesis testing using ANOVA is needed to verify whether the different groups of errors really differ from each other. This is explained by TK Kim and Tae Kyun in their paper [5]. A Post-Hoc test needs to be performed along with ANOVA if the number of groups exceeds two.

Tukey's Test has been explored by Haynes W. in his research paper [6]. Using these techniques, we will create, train and test the effectiveness of our statistical models.

III. PROPOSED MODEL

A. *Null Hypothesis*

Even though the magnitude of overfitting has been reduced, Regression trees still suffer from overfitting even after Pruning. This leads to our following hypothesis.

Hypothesis: Multiple and Lasso Regressions are better at predicting price than the Regression Tree.

B. *Training and Testing Data*

The data is split into training(70% - 563 records) and testing(30% - 241 records) data sets through random sampling (seed was set to 2786).

C. *Lasso Regression*

Using Lasso regression on the training data set, we first select the subset of attributes which lead to optimal/least sum of squared error while predicting the price. It makes use of 10-fold cross-validation to "lasso" the optimal subset of attributes. It uses L1 regularization.

Table – 1: Lasso Regression Summary

LAR Selection Summary

Step	Effect Entered	Number Effects In	CV PRESS
0	Intercept	1	5.47454E10
1	Cylinder_8	2	2.94477E10
2	Make_Cadil	3	2.54198E10
3	Type_Conve	4	1.70491E10
4	Make_SAAB	5	1.0723E10
5	Liter	6	5716511688
6	Model_XLR-V8	7	4455568638
7	Cruise_0	8	4462900633
8	Mileage	9	3141496232
9	Make_Chevr	10	3102016376
10	Model_Corvette	11	2636230790
11	Type_Wagon	12	2434477976
12	Model_STS-V8	13	2241697550
13	Model_Park Ave	14	2022249850
14	Model_9_5	15	2016211162
15	Trim_SS Sedan 4D	16	1870278120
16	Model_STS-V6	17	1767874656
17	Model_Grand Pr	18	1706384400
18	Model_CST-V	19	1594252419
19	Trim_Arc Sedan 4	20	1537014571
20	Trim_Arc Conv 2D	21	1432488055
21	Trim_GT Coupe 2D	22	1357217957
22	Trim_Special Ed	23	1341923945
23	Model_9-2X AWD	24	1207730522
24	Model_Deville	25	1192390216
25	Model_Malibu	26	1140423212
26	Model_Lacrosse	27	1079213317
27	Model_Vibe	28	1057127900
28	Trim_SS Coupe 2D	29	991121705
29	Trim_SVM Hatchba	30	968461697
30	Trim_Sedan 4D	31	906490734
31	Model_Cavalier	32	900413251
32	Model_AVEO	33	895244688
33	Trim_CXS Sedan 4	34	886133715
34	Model_Sunfire	35	866698872
35	Trim_Custom Seda	36	849386379
36	Trim_SVM Sedan 4	37	842938940
37	Model_Grand Am	38	834462359
38	Trim_LS Coupe 2D	39	820399420
39	Trim_LT Coupe 2D	40	809493685
40	Trim_GXP Sedan 4	41	785550309
41	Model_Century	42	780838005
42	Model_L Series	43	757522044
43	Model_G6	44	734553436
44	Trim_GTP Sedan 4	45	709162714
45	Trim_Limited Sed	46	691701586
46	Trim_AWD Sportwa	47	687992677
47	Trim_CXL Sedan 4	48	680838392
48	Trim_DTS Sedan 4	49	674240643
49	Leather_0	50	664002260
50	Trim_Arc Wagon 4	51	662795933
51	Trim_DHS Sedan 4	52	618387221
52	Trim_GT Sportwag	53	612366627
53	Make_Satur	54	609667982
54	Trim_LS Sport Co	55	606909142
55	Model_Classic	56	604819404
56	Trim_SLE Sedan 4	57	601079687
57	Sound_0	58	596479662
58	Trim_GT Sedan 4D	59	597018312
59	Trim_Linear Conv	60	599172036
60	Trim_LT Sedan 4D	61	597976402
61	Trim_Coupe 2D	62	595494703
62	Trim_Conv 2D	63	587687756
63	Trim_LT MAXX Hba	64	586066103
64	Model_9_5 HO	65	585734528
65	Trim_MAXX Hback	66	585894705
66	Trim_LS Sedan 4D	67	586199168
67	Model_Monte Ca	68	582745854*
68	Trim_Quad Coupe	69	583208092
69	Trim_LT Hatchbac	70	583490938
70	Trim_LS Hatchbac	71	585911868
71	Trim_Aero Wagon	72	586112390
72	Trim_LS Sport Se	73	586892601
73	Trim_Aero Sedan	74	587208896

* Optimal Value of Criterion

The LAR Selection summary returns the levels of attributes which need to be chosen to reduce the prediction error.

We can infer from the table-1 that the cross-validated predicted residual error sum of squares (CV PRESS) is the least for the 67 levels of the chosen attributes. Fig. 1 gives us a graphical representation of this. All the chosen 12 attributes, except doors, were lassoed.

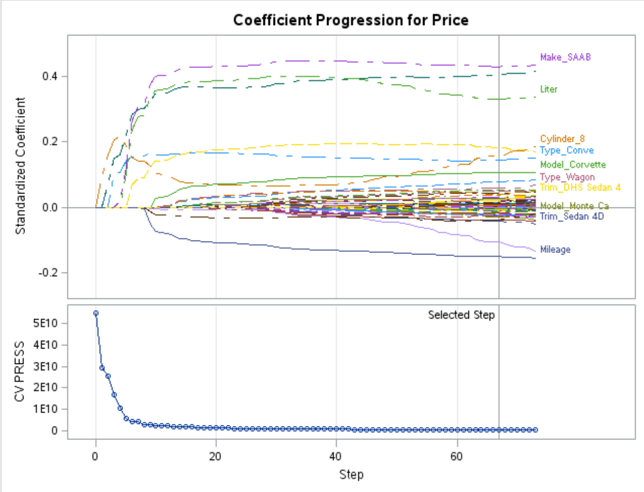

Fig. 1: The coefficients (estimates) of each parameter when other parameters are added is plotted. Also, the CV-Press of the selection process is plotted.

The error rate reaches the minimum value when the above-mentioned levels of the variables were selected for multiple regression. This is shown in Figure 2.

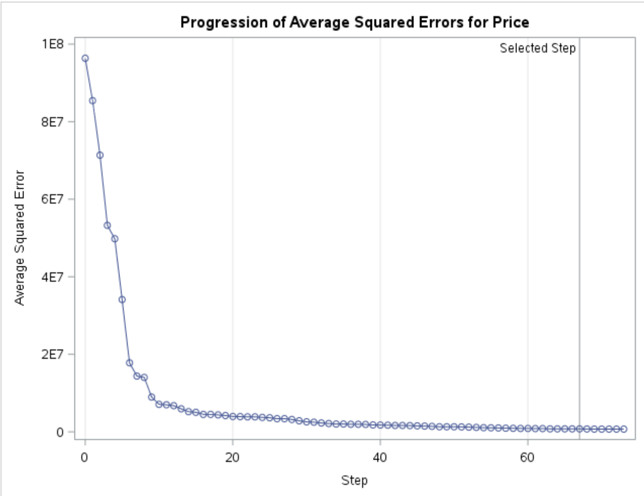

Fig. 2: The Average Square Error is also plotted against the number of levels of variables selected.

Since each level of the categorical variable is treated as a variable on its own (in multiple and Lasso regression), we get 67 estimates.

The prediction model works based on this generated equation:

$$Price = Intercept + P_1{*}E_1 + P_2{*}E_2 + \ldots\ldots + P_{67}{*}E_{67} \quad (1)$$

Where P_1-P_{67} are parameter values while E_1-E_{67} are the parameter estimates. These parameter estimates are tabulated in Table 2.

Table – 2: Parameter Estimates of Lasso Regression

Parameter Estimates		
Parameter	DF	Estimate
Intercept	1	11930
Mileage	1	-0.180375
Model_9-2X AWD	1	-5262.702319
Model_9_5	1	589.004468
Model_9_5 HO	1	83.558352
Model_AVEO	1	-925.178888
Model_CST-V	1	2224.293026
Model_Cavalier	1	-805.439239
Model_Century	1	-450.383125
Model_Classic	1	516.960697
Model_Corvette	1	6499.300407
Model_Deville	1	-6076.484632
Model_G6	1	1524.839816
Model_Grand Am	1	-1417.822185
Model_Grand Pr	1	-1398.133632
Model_L Series	1	-695.355181
Model_Lacrosse	1	1040.680489
Model_Malibu	1	-722.730821
Model_Monte Ca	1	-3.299907
Model_Park Ave	1	4419.996622
Model_STS-V6	1	4615.191143
Model_STS-V8	1	3346.668189
Model_Sunfire	1	-1451.072301
Model_Vibe	1	-450.775368
Model_XLR-V8	1	15127
Trim_AWD Sportwa	1	1105.577065
Trim_Arc Conv 2D	1	3609.398271
Trim_Arc Sedan 4	1	1967.547470
Trim_Arc Wagon 4	1	646.224125
Trim_CXL Sedan 4	1	1167.499627
Trim_CXS Sedan 4	1	2050.191555
Trim_Conv 2D	1	-1044.283584
Trim_Coupe 2D	1	-494.093254
Trim_Custom Seda	1	-813.189144
Trim_DHS Sedan 4	1	2730.946298
Trim_DTS Sedan 4	1	3107.047851
Trim_GT Coupe 2D	1	-1576.659492
Trim_GT Sedan 4D	1	-249.754257
Trim_GT Sportwag	1	864.736852
Trim_GTP Sedan 4	1	1884.237366
Trim_GXP Sedan 4	1	-3057.374358
Trim_LS Coupe 2D	1	-603.093233
Trim_LS Sedan 4D	1	-10.636972
Trim_LS Sport Co	1	-670.111508
Trim_LT Coupe 2D	1	1753.496265
Trim_LT MAXX Hba	1	236.862557
Trim_LT Sedan 4D	1	107.095491
Trim_Limited Sed	1	1674.353968
Trim_Linear Conv	1	345.794386
Trim_MAXX Hback	1	-31.944884
Trim_SLE Sedan 4	1	785.169632
Trim_SS Coupe 2D	1	3154.137720
Trim_SS Sedan 4D	1	5095.123641
Trim_SVM Hatchba	1	-1841.003693
Trim_SVM Sedan 4	1	-937.611706
Trim_Sedan 4D	1	-1056.845083
Trim_Special Ed	1	659.014640
Make_Cadil	1	13886
Make_Chevr	1	-784.202071
Make_SAAB	1	12048
Make_Satur	1	-218.047931
Type_Conve	1	5780.790292
Type_Wagon	1	2014.086276
Cylinder_8	1	4725.325637
Liter	1	2899.110914
Cruise_1	1	52.281909
Sound_0	1	-173.157633
Leather_1	1	221.835813

The parameter estimates of the 67 levels are tabulated here Since Lasso regression heavily relies on the training set to find the best fit levels of attributes, it might miss out on some levels of categorical variables which do not show much association in the training dataset, due to random sampling. This might cause our model to be slightly (maybe even statistically insignificant) underfit, since in-group variance might have been overlooked. Hence, an iterative process is needed to determine the mean error rate.

D. *Multiple Regression*

A general linear model, which models price to the set of selected attributes is trained (on the training data set). The results are tabulated in Table 3. The variables which were selected in Lasso Regression are used here. However here, all the levels of the variables are taken into consideration.

Table – 3: Multiple Regression Summary

The GLM Procedure

Dependent Variable: Price

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	73	53845912247	737615236	910.75	<.0001
Error	489	396039875	809897		
Corrected Total	562	54241952121			

R-Square	Coeff Var	Root MSE	Price Mean
0.992699	4.251916	899.9431	21165.59

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Mileage	1	754087627	754087627	931.09	<.0001
Model	31	51879645703	1673536958	2066.36	<.0001
Trim	38	1201526121	31619108	39.04	<.0001
Make	0	0	.	.	.
Type	0	0	.	.	.
Cylinder	0	0	.	.	.
Liter	0	0	.	.	.
Cruise	1	28758	28758	0.04	0.8506
Sound	1	5915030	5915030	7.30	0.0071
Leather	1	4709008	4709008	5.81	0.0163

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Mileage	1	1159816853	1159816853	1432.05	<.0001
Model	8	493298602	61662325	76.14	<.0001
Trim	31	608810398	19639045	24.25	<.0001
Make	0	0	.	.	.
Type	0	0	.	.	.
Cylinder	0	0	.	.	.
Liter	0	0	.	.	.
Cruise	1	49015	49015	0.06	0.8058
Sound	1	5487410	5487410	6.78	0.0095
Leather	1	4709008	4709008	5.81	0.0163

The result of the GLM procedure with P-value and R^2 values are tabulated along with the type 1 and type 3 error rates.

From this model, we can see that the variable Price and the selected variables are highly correlated since the R-Square (coefficient of determination) value is around 0.9927. This implies that these variables account for about 99.27% of the variance in the Price.

Moreover, both Type 1 and Type 3 SS tables show us that all the variables are significantly correlated with Price (P values < 0.05), except Cruise control, which is confounded when the other variables are held at their mean.
Similar to the GLM Select procedure, this procedure also returns a set of parameter estimates, for numerical variables and every level of the categorical variables.

$$Price = Intercept + P_1{*}E_1 + P_2{*}E_2 + \ldots\ldots + P_n{*}E_n \qquad (2)$$

Where P_1-P_n are parameter values while E_1-E_n are the parameter estimates. These parameter estimates are tabulated in Table 4.

Table – 4: Parameter Estimates of Multiple Regression

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	65357.71411	B	336.7093195	194.11	<.0001
Mileage	-0.18481		0.0048835	-37.84	<.0001
Model 9-2X AWD	-37289.38111	B	614.1831122	-60.71	<.0001
Model 9_3	-35296.16123	B	469.8443603	-75.12	<.0001
Model 9_3 HO	-32915.53020	B	679.1892212	-48.46	<.0001
Model 9_5	-31623.78449	B	451.4965553	-70.04	<.0001
Model 9_5 HO	-31278.61804	B	832.3506337	-37.58	<.0001
Model AVEO	-49587.15746	B	642.7200237	-77.15	<.0001
Model Bonnevil	-40991.14551	B	516.0928169	-79.43	<.0001
Model CST-V	-14670.76189	B	655.6385810	-22.38	<.0001
Model CTS	-30226.42345	B	675.2713617	-44.76	<.0001
Model Cavalier	-47498.91131	B	597.1873761	-79.54	<.0001
Model Century	-43887.65908	B	631.6972266	-69.48	<.0001
Model Classic	-45946.33568	B	640.1245680	-71.78	<.0001
Model Cobalt	-46712.35251	B	601.7780460	-77.62	<.0001
Model Corvette	-23521.01278	B	712.5300647	-33.01	<.0001
Model Deville	-27020.10490	B	641.3631111	-42.13	<.0001
Model G6	-40353.56209	B	617.1102709	-65.39	<.0001
Model GTO	-30411.49383	B	703.3165454	-43.24	<.0001
Model Grand Am	-47085.11233	B	741.4080127	-63.51	<.0001
Model Grand Pr	-42817.18963	B	631.4756193	-67.80	<.0001
Model Impala	-41984.28591	B	603.5127505	-69.57	<.0001
Model Ion	-46263.26174	B	591.6424302	-78.19	<.0001
Model L Series	-45407.97249	B	445.6985609	-101.88	<.0001
Model Lacrosse	-38667.66403	B	490.2095435	-78.88	<.0001
Model Lesabre	-40075.18228	B	515.4683786	-77.75	<.0001
Model Malibu	-43894.77849	B	602.2931867	-72.88	<.0001
Model Monte Ca	-43066.25281	B	726.1099600	-59.31	<.0001
Model Park Ave	-36722.69073	B	454.5671398	-80.79	<.0001
Model STS-V6	-23088.20298	B	654.5603270	-35.27	<.0001
Model STS-V8	-16770.29602	B	629.8807982	-26.62	<.0001
Model Sunfire	-47064.57035	B	703.9717325	-66.86	<.0001
Model Vibe	-46316.43680	B	462.4063161	-100.16	<.0001
Model XLR-V8	0.00000	B	.	.	.
Trim AWD Sportwa	1403.35298	B	452.3758837	3.10	0.0020
Trim Aero Conv 2	3585.16498	B	675.2493313	5.31	<.0001
Trim Aero Sedan	-2691.07765	B	699.8086760	-3.85	0.0001
Trim Aero Wagon	-933.73144	B	842.0462909	-1.11	0.2680
Trim Arc Conv 2D	7337.82741	B	675.2118877	10.87	<.0001
Trim Arc Sedan 4	-77.88681	B	486.3534827	-0.16	0.8728
Trim Arc Wagon 4	782.26079	B	465.9823406	1.68	0.0938
Trim CX Sedan 4D	-2965.34780	B	520.8103669	-5.69	<.0001
Trim CXL Sedan 4	-1497.34465	B	486.7958180	-3.08	0.0022
Trim CXS Sedan 4	0.00000	B	.	.	.
Trim Conv 2D	2855.00108	B	721.6337577	3.96	<.0001
Trim Coupe 2D	-1748.94887	B	550.1784096	-3.18	0.0016
Trim Custom Seda	-2942.25814	B	514.5423072	-5.72	<.0001
Trim DHS Sedan 4	2841.44331	B	664.8961601	4.27	<.0001
Trim DTS Sedan 4	3245.63577	B	715.4464014	4.54	<.0001
Trim GT Coupe 2D	646.26328	B	741.8700293	0.87	0.3841
Trim GT Sedan 4D	-1103.50589	B	562.1036905	-1.96	0.0502
Trim GT Sportwag	1146.16407	B	452.1955261	2.53	0.0116
Trim GTP Sedan 4	1520.55283	B	637.3977100	2.39	0.0174
Trim GXP Sedan 4	2741.49638	B	571.6131934	4.80	<.0001
Trim Hardtop Con	0.00000	B	.	.	.
Trim L300 Sedan	0.00000	B	.	.	.
Trim LS Coupe 2D	-1123.59900	B	550.1303565	-2.04	0.0416
Trim LS Hatchbac	-394.70742	B	662.8277300	-0.60	0.5518
Trim LS MAXX Hba	-567.22900	B	591.6751026	-0.96	0.3382
Trim LS Sedan 4D	-776.01024	B	501.0271130	-1.55	0.1221
Trim LS Sport Co	-1698.76844	B	624.1889147	-2.72	0.0067
Trim LS Sport Se	-809.05535	B	595.6347212	-1.36	0.1750
Trim LT Coupe 2D	2205.96930	B	714.1313029	3.09	0.0021
Trim LT Hatchbac	-369.88643	B	651.4177480	-0.57	0.5704
Trim LT MAXX Hba	-81.94184	B	652.9516792	-0.13	0.9002
Trim LT Sedan 4D	-392.22303	B	530.2066091	-0.74	0.4598
Trim Limited Sed	0.00000	B	.	.	.
Trim Linear Conv	6494.06989	B	529.8423161	12.26	<.0001
Trim Linear Seda	0.00000	B	.	.	.
Trim Linear Wago	0.00000	B	.	.	.
Trim MAXX Hback	-822.49178	B	630.4558787	-1.30	0.1926
Trim Quad Coupe	-441.93382	B	539.9099134	-0.82	0.4135
Trim SE Sedan 4D	-1158.51424	B	545.4710418	-2.12	0.0342
Trim SLE Sedan 4	0.00000	B	.	.	.
Trim SS Coupe 2D	3622.12964	B	723.6504046	5.01	<.0001
Trim SS Sedan 4D	4498.32818	B	630.3936223	7.14	<.0001
Trim SVM Hatchba	-2542.37138	B	620.5971274	-4.10	<.0001
Trim SVM Sedan 4	-1586.36866	B	636.9798038	-2.49	0.0131
Trim Sedan 4D	-1854.33717	B	438.1947081	-4.23	<.0001
Trim Special Ed	0.00000	B	.	.	.
Trim Sportwagon	0.00000	B	.	.	.
Make Buick	0.00000	B	.	.	.
Make Cadil	0.00000	B	.	.	.
Make Chevr	0.00000	B	.	.	.
Make Ponti	0.00000	B	.	.	.
Make SAAB	0.00000	B	.	.	.
Make Satur	0.00000	B	.	.	.
Type Conve	0.00000	B	.	.	.
Type Coupe	0.00000	B	.	.	.
Type Hatch	0.00000	B	.	.	.
Type Sedan	0.00000	B	.	.	.
Type Wagon	0.00000	B	.	.	.
Cylinder 4	0.00000	B	.	.	.
Cylinder 6	0.00000	B	.	.	.
Cylinder 8	0.00000	B	.	.	.
Liter 2	0.00000	B	.	.	.
Liter 3	0.00000	B	.	.	.
Liter 6	0.00000	B	.	.	.
Liter 1.6	0.00000	B	.	.	.
Liter 1.8	0.00000	B	.	.	.
Liter 2.2	0.00000	B	.	.	.
Liter 2.3	0.00000	B	.	.	.
Liter 2.5	0.00000	B	.	.	.
Liter 2.8	0.00000	B	.	.	.
Liter 3.1	0.00000	B	.	.	.
Liter 3.4	0.00000	B	.	.	.
Liter 3.5	0.00000	B	.	.	.
Liter 3.6	0.00000	B	.	.	.
Liter 3.8	0.00000	B	.	.	.
Liter 4.6	0.00000	B	.	.	.
Liter 5.7	0.00000	B	.	.	.
Cruise 0	-30.33805	B	123.3212968	-0.25	0.8058
Cruise 1	0.00000	B	.	.	.
Sound 0	-248.28645	B	95.3859321	-2.60	0.0095
Sound 1	0.00000	B	.	.	.
Leather 0	-261.57621	B	108.4797174	-2.41	0.0163
Leather 1	0.00000	B	.	.	.

The parameter estimates of the 11 selected variables are tabulated here.

The QQ plot for the residual of the price (the difference between the observed and predicted values) and the histogram of the distribution of residuals show us that it approximately follows a normal distribution, with some outliers being present.

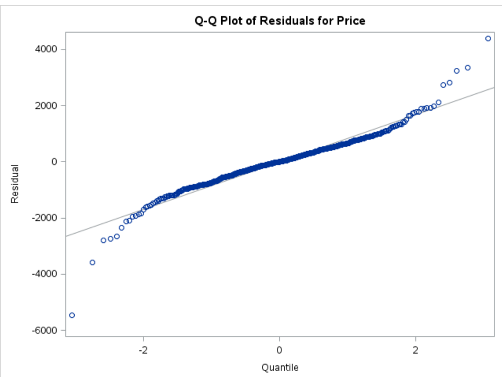

Fig. 3: The QQ-Plot of the residuals are plotted to check for normal distribution

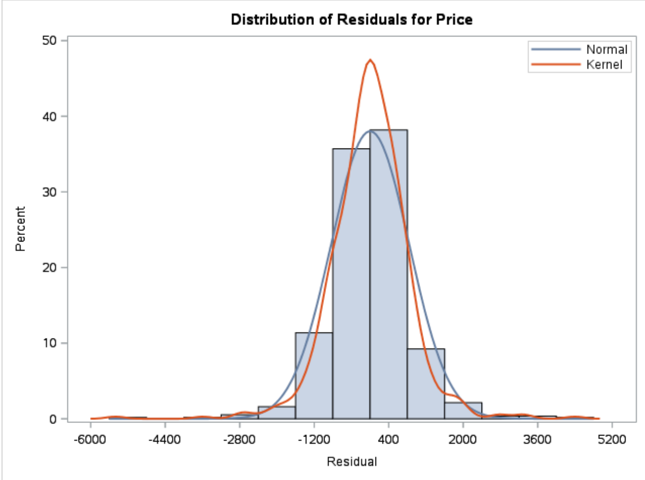

Figure 4: The distribution of residuals is plotted to check for normal distribution.

The studentized residual plot shows us the presence of around 28 outliers in this training data set.

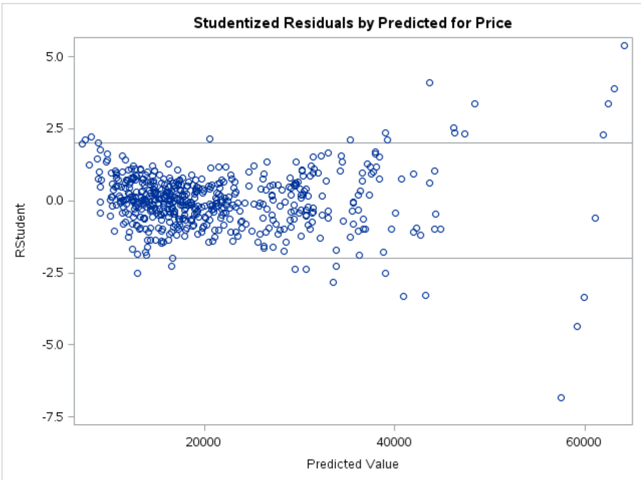

Figure 5: The Studentized residuals are plotted to check for outliers.

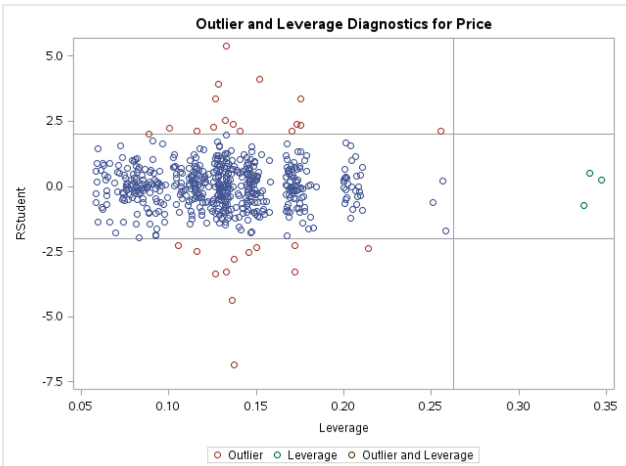

Figure 6: The studentized leverage and outlier plot is used to find whether there are any outliers which heavily influence the prediction model.

The leverage and outlier plots show that these outliers do not hold any leverage. Hence, their absence from the data set doesn't affect the model significantly.

E. *Regression Tree*

A regression tree which models price to the selected subset of attributes is created (by using the training data set) by calling the HPSPLIT Procedure. The results are tabulated in table 5.

Table – 5: HPSPLIT Procedure summary

The HPSPLIT Procedure

Performance Information	
Execution Mode	Single-Machine
Number of Threads	2

Data Access Information			
Data	Engine	Role	Path
WORK.MEX1	V9	Input	On Client

Model Information	
Split Criterion Used	Variance
Pruning Method	Cost-Complexity
Subtree Evaluation Criterion	Cost-Complexity
Number of Branches	2
Maximum Tree Depth Requested	10
Maximum Tree Depth Achieved	10
Tree Depth	10
Number of Leaves Before Pruning	344
Number of Leaves After Pruning	152

Number of Observations Read	563
Number of Observations Used	563

The HPSPLIT Procedure uses Variance for split criteria and Cost-Complexity for Pruning. The number of leaves before and after pruning is also shown.

This tree, before pruning, had 344 leaf nodes. Upon using the cost complexity algorithm for pruning, the number of leaves got reduced to 152. The process of pruning is visually represented in figure 7.

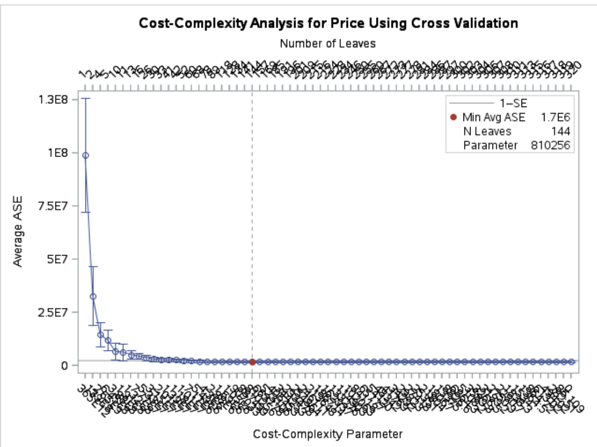

Figure 7: The Average Square Rate is plotted against the number of leaves and cost-complexity parameter to find the minimum ASE.

Here, the minimum average square error is 1.7E6, and that model is selected. The following tree (Fig 8) was produced, which has reduced overfitting. The zoomed-in regression tree (Fig 9) is generated alongside the actual regression tree.

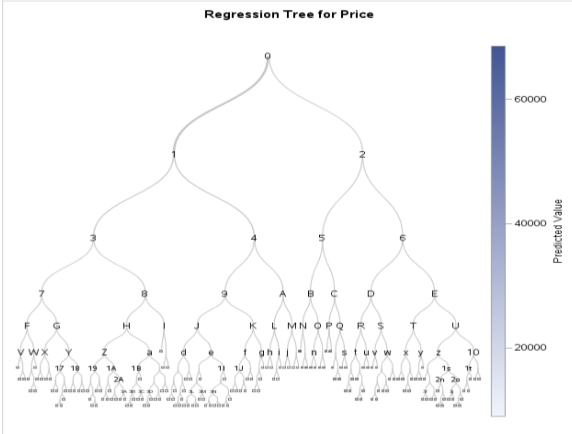

Figure 8: The Regression tree is graphically represented.

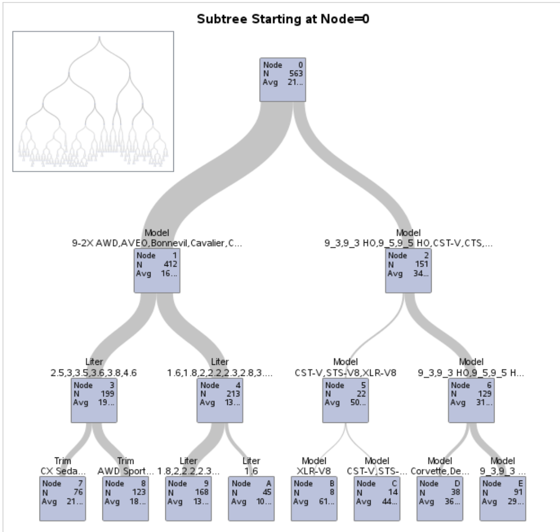

Figure 9: Zoomed in Regression tree.

The order of importance of the variables is also tabulated (table 6). From it, we can infer that the model of the car is most associated with price and that the presence/absence of upgraded sound systems is least associated with price.

Table – 6: HPSPLIT Attribute Importance

Model-Based Fit Statistics for Selected Tree		
N Leaves	ASE	RSS
152	121859	68606340

Variable Importance			
	Training		
Variable	Relative	Importance	Count
Model	1.0000	215640	25
Liter	0.3080	66418.3	5
Trim	0.1855	40004.7	31
Mileage	0.1839	39651.1	80
Type	0.0410	8849.2	3
Leather	0.0106	2296.2	3
Make	0.0076	1629.4	2
Sound	0.0073	1575.6	2

F. *Prediction on test data*

The 3 trained models were used to predict the price of the test data, which contained 241 records. The Observed vs Predicted graphs were plotted for all the three models.

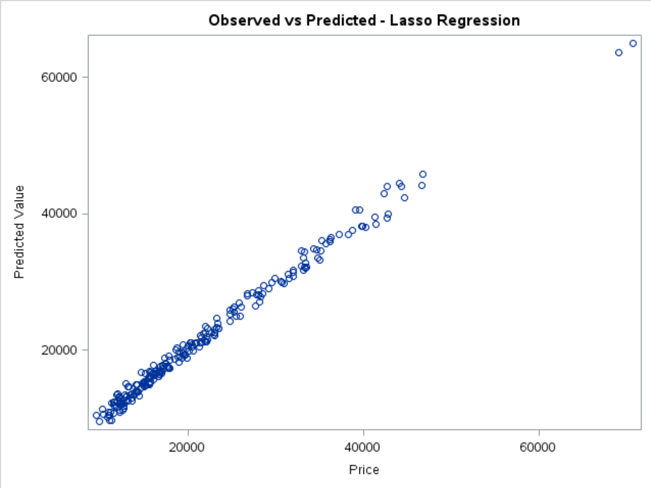

Figure 10: Observed vs Predicted Price – Lasso Regression.

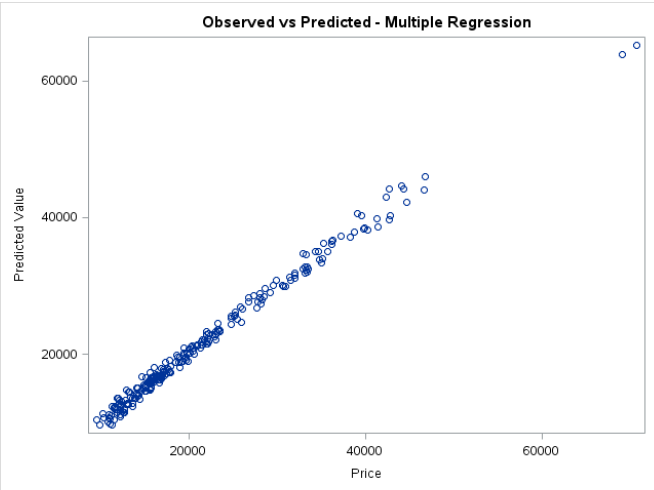

Figure 11: Observed vs Predicted Price – Multiple Regression.

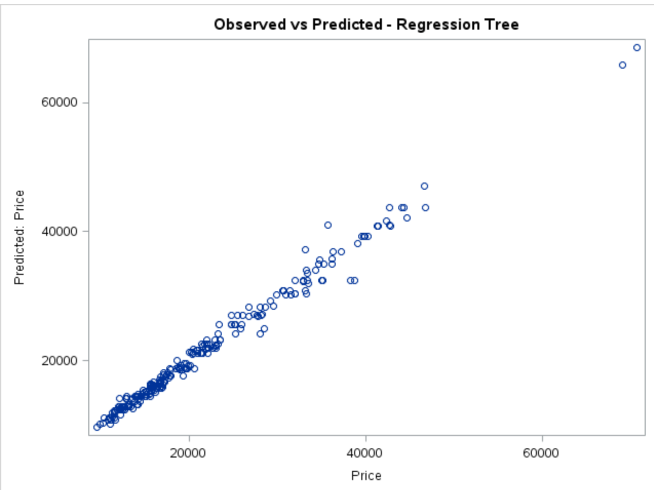

Figure 12: Observed vs Predicted Price – Regression Tree.

The error rates for these models were calculated by using the following formula:

$$Mean\ (\sum (\ |\ (observed - predicted)\ |\ /\ observed\ )\ *\ 100\ ) \quad (3)$$

The results are tabulated below.

Table – 7: Model Error Rates

Model	Error Rate
Lasso Regression	3.581%
Multiple Regression	3.468%
Regression Tree	3.512%

Looking at our models, we see that error rate in multiple regression (3.468%) is smaller than the error rate in Regression tree (3.512%) which is lesser than the error rate in Lasso Regression (3.581%).

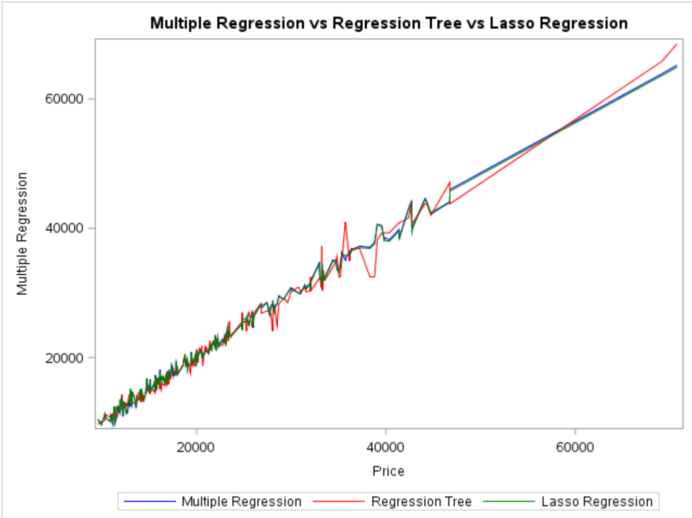

Figure 13: A color-coded line graph comparing the price predictions of the different models with the actual value

However, from this, we can't conclude that our hypothesis holds good since the error rates were found only on one variation of the training and testing data set. By iterating this process[8](with the selection of different records by varying the seed of the random sampling procedure), we will get a set of error rates of lasso regression, multiple regression and regression tree, for the same variation of the data set.

G. *Iterative ANOVA based comparison of models*

Using One-way Analysis Of Variance (ANOVA) we need to verify whether the error rates of these models differ significantly from each other.
The process was run 35 times, and the error rates for lasso regression, multiple regression, and regression tree were noted (Table 8) along with the respective seeds of splitting for reproducibility.

Retrieval Number: A10421291S319/2019@BEIESP
DOI:10.35940/ijeat.A1042.1291S319

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Table – 8: Seed – Error Matrix

Total rows: 35 Total columns: 4 Rows 1-35

Seed	multiple	tree	lasso
2786	3.46839529	3.51176948	3.58138216
1589	3.6757781	3.5680421	3.57764978
100	3.67042	4.14094084	4.01717736
1458	3.49021115	3.24932075	3.48042261
2607	3.25016677	3.91970384	3.28921334
8457	3.61699478	4.40631117	3.71202329
5841	3.44307372	3.67629277	3.6288628
6985	3.68199985	3.77225492	3.64147977
4185	3.51752289	3.58014847	3.45390467
1208	3.58700681	4.02278469	3.48855108
7408	3.55941912	3.48832635	3.56545587
7985	3.38059236	3.7056248	3.33113402
27	3.32998887	3.41531015	3.24091976
3451	3.64931838	3.89251644	3.71782049
8	3.6142105	4.21475854	3.71334439
587	3.0037316	3.65458366	3.10518136

This table contains the error rates of the three models for 35 different variations of training and test data.

The data were recoded to perform ANOVA.

Table – 9: Recoded Error Data

Total rows: 105 Total columns: 2 Rows 1-100

	error	type
58	4.0627354	tree
59	3.94371457	tree
60	3.85089763	tree
61	3.591572	tree
62	3.63890433	tree
63	3.58138216	lasso
64	3.57764978	lasso
65	4.01717736	lasso
66	3.48042261	lasso
67	3.28921334	lasso
68	3.71202329	lasso
69	3.6288628	lasso
70	3.64147977	lasso
71	3.45390467	lasso
72	3.48855108	lasso
73	3.56545587	lasso

This table contains the recoded error rates.

The ANOVA procedure was carried out, and the results were tabulated in table 10.

Table – 10: ANOVA Summary

The ANOVA Procedure

Class Level Information

Class	Levels	Values
type	3	lasso multiple tree

Number of Observations Read	105
Number of Observations Used	105

The ANOVA Procedure

Dependent Variable: error error

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1.89535222	0.94767611	20.74	<.0001
Error	102	4.66027047	0.04568893		
Corrected Total	104	6.55562270			

R-Square	Coeff Var	Root MSE	error Mean
0.289119	5.961876	0.213750	3.585275

Source	DF	Anova SS	Mean Square	F Value	Pr > F
type	2	1.89535222	0.94767611	20.74	<.0001

With the P-value being lesser than 0.05, we can confirm that the error rates are significantly different from each other. Their distribution is also plotted (Fig 14).

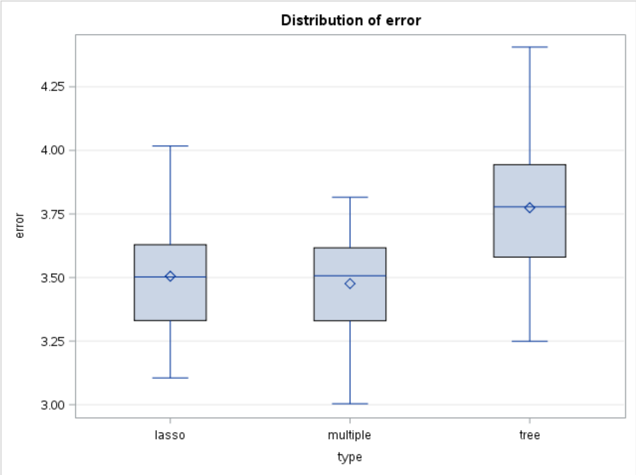

Figure 14: The distribution of the error rates of each model type is represented using box plots.

Table – 11: Mean Error Rates

Level of type	N	error	
		Mean	Std Dev
lasso	35	3.50528606	0.19866047
multiple	35	3.47601284	0.18968152
tree	35	3.77452740	0.24825250

The mean error rates of the models (Table 11) might be misleading, since we can't be sure about which groups/models have significantly different means from the other. This is due to the existence of more than 2 groups/levels. One-Way ANOVA can only find out whether there exists any significant difference between any of the groups. To get a clearer picture, we need to perform a post-hoc test to find the groups which have significantly different means.

We are performing a Tukey's test (Tukey's Honest Significant Difference Test) to find out the groups which are actually different from each other.

The test compares all possible pairs of means and checks for statistically significant differences between them. Since the sample size for all the groups is the same, we do not use the Tukey-Kramer Method[7], and use the standard version of the algorithm.

Table – 12: Tukey's Test Summary

Alpha	0.05
Error Degrees of Freedom	102
Error Mean Square	0.045689
Critical Value of Studentized Range	3.36358
Minimum Significant Difference	0.1215

From table 12 we can get the critical value of the studentized range and the minimum significant difference. The result is plotted graphically (Fig 15).

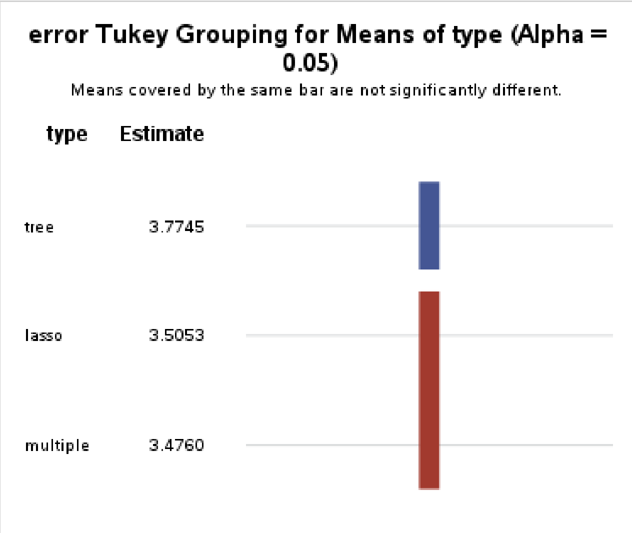

Figure 15: The result of Tukey's Test is plotted in such a way that the means which are covered by the same bar are not significantly different.

From Fig. 15, we can infer that the mean error rates of lasso regression models and multiple regression models are not significantly different, but the mean error rate of regression trees are higher and significantly different from the other two.

IV. CONCLUSION AND FUTURE ENHANCEMENT

The prediction error rate of all the models was well under the accepted 5% of error. But, on further analysis, the mean error of the regression tree model was found to be more than the mean error rate of the multiple regression and lasso regression models. Even though for some seeds the regression tree has better accuracy, its error rates are higher for the rest. This has been confirmed by performing an ANOVA. Also, the post-hoc test revealed that the error rates in multiple regression models and lasso regression models aren't significantly different from each other. To get even more accurate models, we can also choose more advanced machine learning algorithms such as random forests, an ensemble learning algorithm which creates multiple decision/regression trees, which brings down overfitting massively or Boosting, which tries to bias the overall model by weighing in the favor of good performers. More data from newer websites and different countries can also be scraped and this data can be used to retrain these models to check for reproducibility.

REFERENCES

[1] Shonda Kuiper (2008) Introduction to Multiple Regression: How Much Is Your Car Worth?, Journal of Statistics Education, 16:3, DOI: 10.1080/10691898.2008.11889579

[2] Geurts P. (2009) Bias vs Variance Decomposition for Regression and Classification. In: Maimon O., Rokach L. (eds) Data Mining and Knowledge Discovery Handbook. Springer, Boston, MA

[3] Robert T. (1996) Regression Shrinkage and Selection Via the Lasso. In: Journal of the Royal Statistical Society: Series B (Methodological) Volume 58, Issue 1

[4] Hastie, Trevor, and Daryl Pregibon. Shrinking trees. AT & T Bell Laboratories, 1990.

[5] Kim, Tae Kyun. "Understanding one-way ANOVA using conceptual figures." Korean journal of anesthesiology 70.1 (2017): 22.

[6] Haynes W. (2013) Tukey's Test. In: Dubitzky W., Wolkenhauer O., Cho KH., Yokota H. (eds) Encyclopedia of Systems Biology. Springer, New York, NY

[7] Jaccard, James, Michael A. Becker, and Gregory Wood. "Pairwise multiple comparison procedures: A review." Psychological Bulletin 96.3 (1984): 589.

[8] Dupac, Václac, ed. Sampling from a finite population. Marcel Dekker, Incorporated, 1981.

AUTHORS PROFILE

Pattabiraman Venkatasubbu obtained his Ph.D. from Bharathiar University, India. He has a total Professional experience of 19 years working in various prestigious institutions. He is currently a Professor at Vellore Institute of Technology, Chennai Campus, India. He has authored several books in the field of Computer Science. He is a Senior member of International Association of Computer Science and Information Technology (IACSIT) also he is member in various professional societies namely ACM, IEEE, ISTE, CSI, Society for Research in Information Security and Privacy- SRISP and Academy & Industry Research Collaboration Center (AIRCC). Dr. Pattabiraman's teaching and research expertise covers a wide range of subject area including Knowledge discovery and Data mining, Big Data Analytics, Machine Learning, Deep Learning, Database technologies, Data Structures and Analysis of Algorithms etc., He has also received several awards in his career.

Mukkesh Ganesh is a B.tech Computer Science student at Vellore Institute of Technology, Chennai. A budding ML and AI enthusiast, who is working on leveraging the power of AI to solve a variety of problems. He is currently researching on the detection of anomalies using deep learning methods and time series forecasting of kidnapping rates in India. He is also interested in the prospect of enhancing Edge Computing in the field of IOT, to decentralize the heavy server-level processing. His other areas of interest include algorithm design, parallel and distributed computing and theory of computation. He is also working on several projects which span across different fields in Computer Science. He is a finalist of Smart India Hackathon 2019 software edition, and was one of the youngest team leaders in the government organized competition.