ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 9(44), DOI: 10.17485/ijst/2016/v9i44/105078, November 2016* ISSN (Online) : 0974-5645

Automobile Engine Performance Analysis using Regression Technique

Kartik Sharma*, A. Sai Sabitha and Abhay Bansal

Department Computer Science and Engineering, Amity University, Noida - 201313, Uttar Pradesh, India;
kartiksharma131@gmail.com, saisabitha@gmail.com, abhaybansal@hotmail.com

Abstract

The need for automobile is growing day by day and researches have been carried out in different fields of automobile industry like design, manufacturing, sales, production etc. In the field of engine design, it is required to take into consideration all the parameters affecting the design and performance of it. There are many parameters that plays vital role in engine design and there is need to determine the measures for improving the engine's performance. The data mining technique can be used to determine these performance measures. The experimental study is conducted using multi-linear regression for various dependent and independent variables using a data mining tool. Principal Component Analysis (PCA) is used to emphasis the selection of attributes and to identify the patterns in automobile dataset.

Keywords: Automobiles, Data Mining, Linear Regression, Multi-Linear Regression, Prediction Techniques

1. Introduction

Automobile industry is the corporate world of manufacturing, marketing and trading self-powered vehicles, including passenger cars, sports car, buses and trucks, and other commercial and utility vehicles. The industry has become the requirement of this fast growing world and is the backbone of the business world. This sector is rapidly expanding day by day due to which there is large need of automobiles as many foreign auto companies are showing their full interest to setup their Auto Industries in various parts of India[1]. The buying power of people has increased with the rise in incomes and emerging middle class[2]. The demands for automobiles have increased manifold due to the enhanced infrastructure and ease of access of commercial automobiles for distant markets[3]. Extensive research works have been carried out in this industry to improve features like design, manufacturing[4], production and performance. The performance issues can be related to engine displacement, acceleration, fuel efficiency etc. The purpose of analysis is to understand the vital parameters that affect the stability and performance of an automobile.

Data Mining techniques like clustering, classification, regression etc. Are widely used to understand the relationship between parameters mentioned above. The regression model is used to identify the important parameters related to performance. The paper is structured as follows: Theoretical Background, Methodology, Experimental Setup, Case Study and Conclusion.

2. Theoretical Background

The vast automobile industry research activities include Engines, Power train, Design, Quality, Modelling, Simulation, and Manufacturing[5,6].

The application of Data Mining can be seen in various fields of Automobile sector which includes manufacturing[4], production[7], performance, safety etc[8]. Data Mining techniques such as clustering, fuzzing[9], classification algorithm, regression algorithm etc helps to find patterns or trends in large data set. The table below depicts a data mining methods used in different applications of automobile industry (Table1).

2.1 Prediction Technique

Regression analysis is a geometric process for studying and assessing the interactions between variables. It comprises of several methods which are used for demonstrating

* *Author for correspondence*

Table 1. Data mining analysis in automobile industry

Year	Name of the Author(S)	Application	Data Mining Methods
2009	Qian Zhou	Vehicle Report-Stop	Data Mining & Data Warehousing[10]
2010	Rudolf Kruse, Matthias Steinbrecher and Christian Moewes	Car lifecycle stages	Clustering[11]
2014	M.Jayakameswaraiah and S.Ramakrishna	Car Manufacturing	ID3Algorithm[12]
2011	Hanumanthappa	No. of cars Manufactured	Linear regression[13]
2011	S.Gunasekaran and C.Chandraskaran	Automobile industry data	Clustering[14]
2012	Marco Hulsman, Christoph M. Freidrich and Dirk reith	Sales forecast	Time series analysis and Classical algorithm[15]
2009	Liu Gaojun and Long Boxue	Sales forecast	Math statistics and neural networks[16]

and examining numerous variables, when the emphasis is being laid on the association between a dependent variable and one or more "predictors" or independent variables.

2.1.1 Types of Regression

Simple Linear Regression: A linear regression is a method which is used to show the association among the independent variables or predictors and a single dependent variable.

$y = w_{0+} w_1 x$, where w_o and w_1 are regression coefficients.

Multi-Linear Regression: The term multi-linear regression is a type of linear regression which is comprising of more than two or two independent variables.

$y = w_0 + w_1 x_1 + w_2 x_2$, where w_0, w_1 and w_2 are regression coefficients.

2.2 Validity Measures

2.2.1 ANOVA Table and the Coefficient of Determination R^2

The total of the dependent variables' sum of squares (SCT) is in the form of sum of squares which is clarified by the model is partitioned by ANOVA table and the residual sum of squares (SCE) is not clarified by the model. The coefficient of determination is given by the ratio of SCE and SCT (Equation (1)).

$$R^2 = 1-\left(\frac{SCR}{SCT}\right) * \left(\frac{n-1}{n-p-1}\right) \quad (1)$$

2.2.1 T-TEST

The next step is the assessment of the independent variables' the influence in model. For each coefficient related with an independent variable, we test the null hypothesis (Equation (2)):
The statistical test is given as

$$t_j = \frac{a^j}{\sigma_{a_j}} \quad (2)$$

$\sigma\hat{}_{a\hat{}j}$ is the standard error of the estimated coefficient. The diagonal of the covariance matrix of the estimated coefficients provides its squared value (Equation (3)).

$$\Omega = \sigma_e^2 (X'X)^{-1} \quad (3)$$

σ_e^2 is the squared of the standard error of regression (Equation (4)). The regression's standard error is obtained with the following formula given below:

$$\sigma_e = \sqrt{\frac{SCR}{n-p-1}} \quad (4)$$

2.2.2 F-TEST

F-test is used to test the significance of model as a whole. Here F stands for fisher distribution (Equation. (5)). With degrees of freedom (p,n-p-1). The model is highly significant if it has lower p-value and vice-versa.

$$F = \frac{SCE/p}{SCR/n-p-1} \quad (5)$$

2.2.3 Principal Component Analysis (PCA)

Principal component technique is used to show variation and find strong pattern within a dataset. PCA combines the essence of attributes by creating an alternative, smaller

set of variables. It is a technique used for eliminating the dimensions by projecting original data into smaller space, resulting in dimensionality reduction.

3. Methodology

The collection of dataset was manually done from various automobile websites. The missing values (Noise) was removed and the clean data was used for selection of attributes. PC analysis was done and regression technique was applied using data mining tool. The results are evaluated and obtained output is analyzed. Figure 1 shows the flowchart of methodology adopted for the proposed analysis.

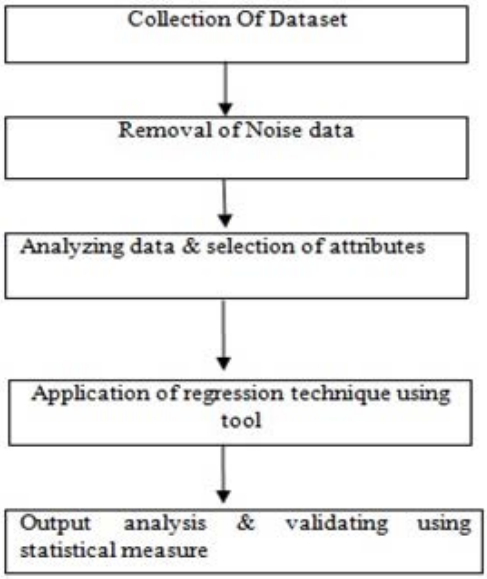

Figure 1. Methodology of proposed analysis.

NAME OF AUTOMOBILE	Engine Displacement (in cc)	Maximum Power (bhp)	Maximum Torque (Nm)	No. of Cylinders
AUDI A3	2000	143	320	4
ACURA ILX	2400	201	180	4
ACURA MDX	3500	290	267	6
ACURA NSX	3500	573	550	6
ACURA RDX	3500	279	252	6
ACURA RL	3700	300	271	6
ACURA RLX	3500	377	314	6
ACURA TL	3500	280	254	6
ACURA TLX	2400	209	182	4
ACURA TSX	2400	201	170	4
ACURA ZDX	3700	300	270	6
ALFA-ROMEO 4C SPIDER	1750	240	258	4
ALFA-ROMEO GIULIETTA	1400	150	175	4
ALFA-ROMEO MiTo	1400	140	170	4
ALFA-ROMEO MiTo RACER	900	105	106	4
ASTON MARTIN DB9	6000	470	600	12
ASTON MARTIN RAPIDE	6000	470	600	12

Figure 2. Automobile dataset.

4. Experimental Setup

4.1 Dataset

The dataset of various automobiles was collected from internet sources[17–23]. The attributes for collected are as follows: Engine Displacement, Torque, Power, Fuel Efficiency, Acceleration, Top Speed and Price. For performance analysis, three prime attributes (Engine Displacement, Maximum Torque generated and Maximum Power delivered) were chosen based on PCA (Principal Component Analysis). Figure 2 shows the screenshot of Automobile dataset.

4.2 Principal Component Analysis

PCA is used for eliminating dimensions by projecting original data into smaller space. Figure 3 shows screenshot of PCA output analysis for dimensionality reduction.

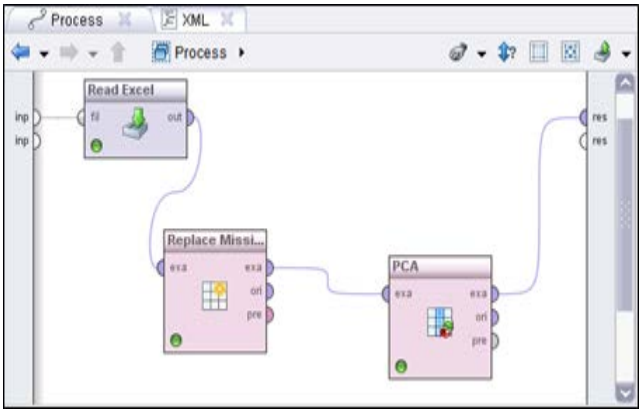

Figure 3. Screenshot of PCA output analysis.

The information about the individual and cumulative contribution of principal component to the data variance is obtained from the Eigen values[24]. The selection of principal component is carried out using this Eigen values. Figure 4 shows the Eigen values table of PCA.

Component	SD	Proportion of variance	Cumulative
PC 1	1663.245	0.99	0.99
PC 2	140.129	0.007	0.997
PC 3	86.305	0.003	1
PC 4	23.364	0	1
PC 5	3.533	0	1
PC 6	2.499	0	1
PC 7	0.866	0	1

Figure 4. Eigen values table.

The variance threshold is set to 95% and PC1, PC2 and PC3 are to be considered for selection of parameters as it clearly explains 99% of the variance. Figure 5 shows the eigen vector of principal components.

	pc1	pc2	pc3	pc4	pc5	pc6	pc7
Top Speed (kmph)	0.027	0.088	-0.289	-0.951	-0.002	0.064	-0.006
Acceleration (0-100 kmph)	-0.002	-0.006	0.02	0.06	0.046	0.997	-0.021
Engine Displacement (in cc)	0.986	-0.153	0.064	-0.006	0.001	0	0.002
Maximum Power (bhp)	0.109	0.319	-0.891	0.304	0.012	0.001	0.005
Maximum Torque (Nm)	0.122	0.931	0.343	-0.014	-0.001	0	-0.003
Fuel Economy (kmpl)	-0.002	-0.002	0.01	-0.009	0.998	-0.044	0.05
No. of Cylinders	0.001	-0.002	-0.004	0.006	0.049	-0.023	-0.999

Figure 5. Eigen vector table.

The parameters considered for the analysis are as follows - Engine displacement, maximum torque, and maximum power as they have highest eigen values from selected principal components. To understand the relationship between independent Vs dependent variables and for analysis of the data, Tanagra a data mining tool was used.

4.3 Tanagra

Tanagra is an "open source project" and free data mining tool for academic and research purposes. It proposes several data mining methods from exploratory data analysis, statistical learning and machine learning.

"Define status" of Tanagra helps to set input attributes (independent variables) and the target value (dependent variable). Figure 6 shows the screenshot of "Define Status" tool.

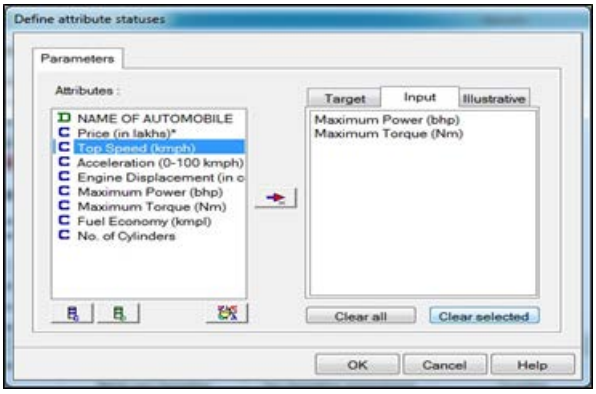

Figure 6. Screenshot of "define status" tool.

For the analysis, Multi-Linear regression tool was considered, Engine Displacement was chosen as dependent variable and Maximum Torque and Maximum Power as independent variables. Figure 7 shows the screenshot of application of Multi-Linear Regression.

Figure 7. Screenshot of multi-linear regression.

5. Analysis

Based on principal component analysis the features chosen are: Engine Displacement as independent variable and, Maximum Torque and Maximum Power as independent variables for Multi-Linear regression analysis on data mining tool.

The result of the regression analysis is shown in Figure 8.

Endogenous attribute	Engine Displacement (in cc)
Examples	320
R^2	0.824772
Adjusted-R^2	0.823667
Sigma error	689.531484
F-Test (2,317)	746.0375 (0.000000)

Figure 8. Results table coefficient of determination.

5.1 R^2- Coefficient of Determination

The coefficient of determination (R^2) is a key output of regression analysis. It is inferred as the proportion of the variance in the dependent variable that is predictable from the independent variable. Its value ranges from 0 to 1.

- The value of $R^2 = 0$ means that the dependent variable cannot be predicted from the independent variable.
- The value of $R^2 = 1$ means that the dependent variable cannot be predicted without error from the independent variable.
- The value of R^2 between 0 and 1 indicates the extent to which the dependent variable is predictable. An R^2 of 0.40 means that 40% of the variance in component Y is predictable from component X.

The value of R^2 observed is 0.824 and it implies that 82.4% of the variance in Engine displacement is predictable from Max. Torque generated and Max. Power delivered by an automobile.

5.2 F-Test Analysis

The significance of the model is evaluated using p-value generated from the F-test (Equation (5)). The range of F-test varies from zero to an arbitrarily large number. The F-test value was found out to be 746.03.

5.2.1 P-Value Analysis

P-values are the probability of obtaining an effect at least as extreme as the one in the sample data, assuming the truth of the null hypothesis. If a p-value is less than or equal to the significance level of $\alpha = 0.05$ (or5%), then the model is considered as significant.

The p-value was found out to be 0.00. Figure 9 shows the ANOVA table.

Analysis of variance

Source	xSS	d.f.	xMS	F	p-value
Regression	709412530.2574	2	354706265.1287	746.0375	0.0000
Residual	150718812.7301	317	475453.6679		
Total	860131342.9875	319			

Figure 9. ANOVA table.

5.3 Residual Analysis

The difference between the observed value of the dependent variable (y) and the predicted value ($\hat{y}$) is called the residual (e). Each data point has one residual.

Residual = Observed value - Predicted value

$e = y - \hat{y}$

Figure 10 shows the residual analysis of the dataset.

NAME OF AUTOMOBILE	Engine Displacement	Maximum Power	Maximum Torque	Fuel Economy	Pred_lmreg_1	Err_Pred_lmreg_1
AUDI A3	2000	143	320	20.38	1999.44	0.555823
ACURA ILX	2400	201	180	15.3	2033.85	366.148
ACURA MDX	3500	290	267	12	2677.81	822.186
ACURA NSX	3500	573	550	5	4737.93	-1237.93
ACURA RDX	3500	279	252	10.63	2589.9	910.095
ACURA RL	3700	300	271	10	2738.86	961.142
ACURA RLX	3500	377	314	13.6	3232.79	267.21
ACURA TL	3500	280	254	12.33	2599.14	900.857
ACURA TLX	2400	209	182	15	2080.34	319.663
ACURA TSX	2400	201	170	10.6	2014.27	385.734
ACURA ZDX	3700	300	270	9.78	2736.9	963.1
ALFA-ROMEO 4C SPIDER	1750	240	258	20	2394.14	-644.14
ALFA-ROMEO GIULIETTA	1400	150	175	17.8	1752.69	-352.691
ALFA-ROMEO MiTo	1400	140	170	18.51	1689.69	-289.689
ALFA-ROMEO MiTo RACER	900	105	106	18.51	1378.1	-478.105
ASTON MARTIN DB9	6000	470	600	9.8	4287.8	1712.2
ASTON MARTIN RAPIDE	6000	470	600	10.9	4287.8	1712.2
ASTON MARTIN VANQUISH	6000	565	620	8	4832.46	1167.54
ASTON MARTIN VANTAGE	6000	510	570	8.5	4441.88	1558.12
ASTON MARTIN ZAGATO	6000	510	570	8	4441.88	1558.12
AUDI A3 CABRILIOT	1800	177	250	16.6	2043.25	-243.253

Figure 10. Residual analysis of dataset.

5.4 Case Study-1

The parameters chosen for the analysis are described as: Torque, Power and Fuel Economy as explanatory variables whereas Engine Displacement as response variable.

The value of coefficient of determination (R^2) was found out to be 0.839 which means that 83.9% of variance in Engine displacement is predictable from Torque, Power and Fuel Economy. Figure 11 shows the table of coefficient of determination R^2.

Endogenous attribute	Engine Displacement (in cc)
Examples	320
R^2	0.839620
Adjusted-R^2	0.838098
Sigma error	660.713873
F-Test (3,316)	551.4417 (0.000000)

Figure 11. Results table coefficient of determination.

The table describes the analysis of variance for the new chosen parameters. Figure 12 shows the ANOVA table.

Analysis of variance

Source	xSS	d.f.	xMS	F	p-value
Regression	722183811.2424	3	240727937.0808	551.4417	0.0000
Residual	137947531.7451	316	436542.8220		
Total	860131342.9875	319			

Figure 12. ANOVA table.

5.5 Case Study-2

The parameters chosen for the second analysis are described as: Torque, Power and No. of cylinders as explanatory variables whereas Engine Displacement as response variable.

Endogenous attribute	Engine Displacement (in cc)
Examples	320
R^2	0.924447
Adjusted-R^2	0.923730
Sigma error	453.486684
F-Test (3,316)	1288.8325 (0.000000)

Figure 13. Results table coefficient of determination.

The value of coefficient of determination (R^2) was found out to be 0.924 which means that 92.4% of variance in Engine displacement is predictable from Torque, Power and No. of cylinders. Figure 13 shows the coefficient of determination for this case.

The table describes the analysis of variance for the new chosen parameters. Figure 14 describes the ANOVA table for this case.

Analysis of variance

Source	xSS	d.f.	xMS	F	p-value
Regression	795145888.4851	3	265048629.4950	1288.8325	0.0000
Residual	64985454.5024	316	205650.1725		
Total	860131342.9875	319			

Figure 14. ANOVA table.

5.5 Case Study-3

The parameters chosen for the third analysis are described as: Torque, Power and Top Speed as explanatory variables whereas Engine Displacement as response variable.

The value of coefficient of determination (R^2) was found out to be 0.827 which means that 82.7% of variance in Engine displacement is predictable from Torque, Power and Top Speed. Figure 15 describes the coefficient of determination for this case.

Endogenous attribute	Engine Displacement (in cc)
Examples	320
R^2	0.827476
Adjusted-R^2	0.825838
Sigma error	685.272784
F-Test (3,316)	505.2101 (0.000000)

Figure 15. Results table coefficient of determination.

The Table describes the analysis of variance for the new chosen parameters. Figure 16 shows the ANOVA table for this case.

Analysis of variance

Source	xSS	d.f.	xMS	F	p-value
Regression	711738125.7841	3	237246041.9280	505.2101	0.0000
Residual	148393217.2034	316	469598.7886		
Total	860131342.9875	319			

Figure 16. ANOVA table.

The variation of Max Torque generated with Engine

Displacement of an automobile is depicted through a scatter plot. Figure 17 shows the scatter plot of Maximum Torque Vs engine displacement.

The variation of Max Power delivered with Engine Displacement of an automobile is depicted through a scatter plot. Figure 18 shows the scatter plot of Maximum Power Vs Engine Displacement.

6. Acknowledgement

The authors wish to acknowledge various automobile websites supporting this research by providing important information in collection of dataset.

7. Conclusion

The regression analysis shows that for the response variable- Engine Displacement, the identified independent variables are: Maximum Torque and Maximum Power, Since the R^2 values for all the case studies were greater than 80%. Statistical measures were used for analysis and validation.

The research work can be done for performance issues associated with other parameters like acceleration, no. of

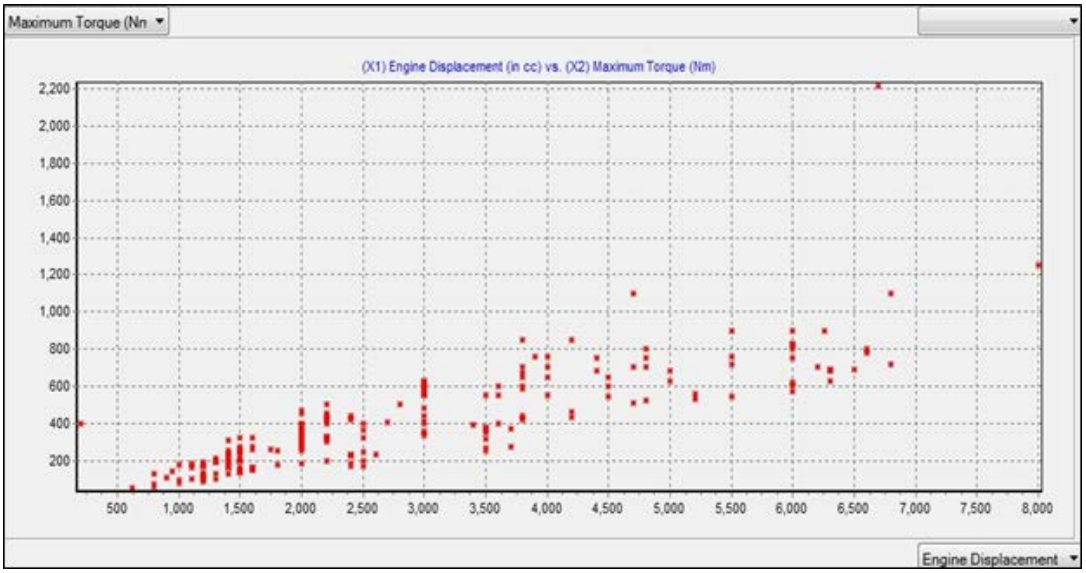

Figure 17. Torque vs engine displacement.

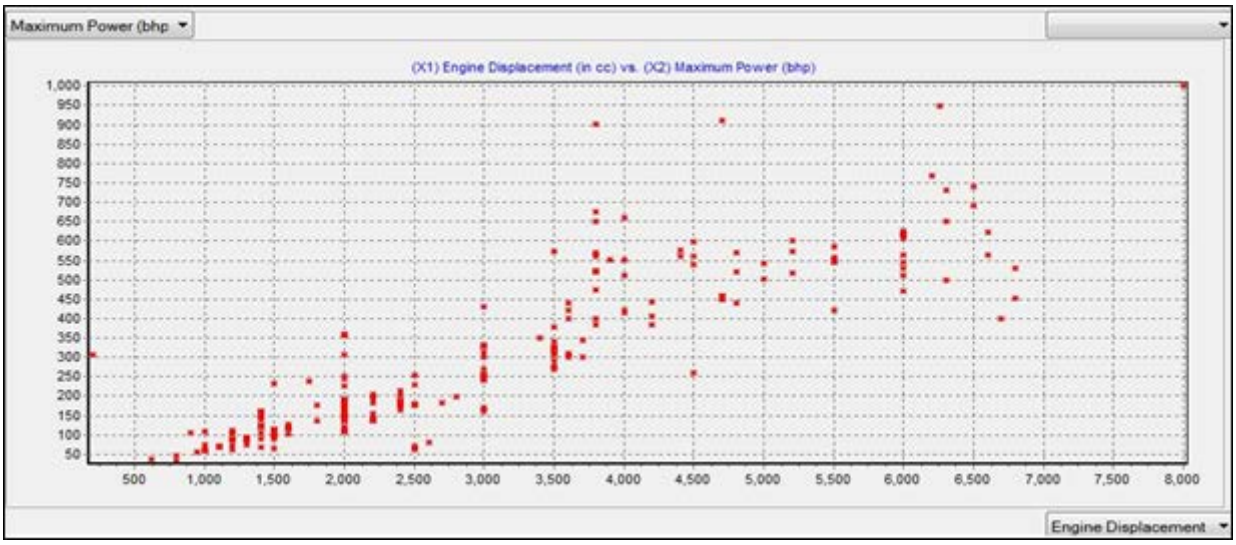

Figure 18. Max. power vs engine displacement.

cylinders, top speed etc. The work can be further extended for predicting a car's price using other features or engine specifications of an automobile. This problem can be further enhanced using other regression techniques and evaluated using evaluation models like cross validation.

7. References

1. Hülsmann M, Borscheid D, Friedrich CM, Reith D. General sales forecast models for automobile markets based on time series analysis and data mining techniques. Proceedings of the 11th International Conference on Advances in Data Mining:Applications and Theoretical Aspects; Springer Berlin Heidelberg. 2011 Aug 30. p. 255–69.
2. Gaojun L, Boxue L. The research on combination forecasting model of the automobile sales forecasting system. International Forum on Computer Science-Technology and Applications; 2009 Dec. p. 3.
3. Gunasekaran S, Chandrasekaran C. A survey on automobile industries using data mining techniques. International Journal of Science and Advanced Technology. 2011; 1(4):30–5.
4. Sa-ngasoongsong A, Bukkapatnam ST, Kim J, Iyer PS, Suresh RP. Multi-step sales forecasting in automotive industry based on structural relationship identification. International Journal of Production Economics. 2012 Dec 31; 140(2):875–87.
5. Dai Q, Zhong R, Huang GQ, Qu T, Zhang T, Luo TY. Radio frequency identification-enabled real-time manufacturing execution system: A case study in an automotive part manufacturer. International Journal of Computer Integrated Manufacturing. 2012 Jan 1; 25(1):51–65.
6. Sambhe RU, Dalu RS. Six sigma implementation in Indian medium scale automotive enterprises- A review and agenda for future research. International Journal of Six Sigma and Competitive Advantage. 2011 Jan 1; 6(3):224–42.
7. Blázquez L, González-Díaz B. International automotive production networks: How the web comes together. Journal of Economic Interaction and Coordination. 2016 Apr 1; 11(1):119–50.
8. Shende V. Analysis of research in consumer behavior of automobile passenger car customer. International Journal of Scientific and Research Publications. 2014 Feb; 4(2):1.
9. Zhou Q, Liu YS. Applying research of data mining technology on the analysis of vehicles report-stop fraud. 1st International Workshop on Database Technology and Applications; 2009 Apr 25. p. 315–8.
10. Kumar V, Devendra V. Fuzzy procedure for the selection of car among various brands. International Journal of Engineering. 2013; 6(3):337–42.
11. Jayakameswaraiah M, Ramakrishna S. Design and development of data mining system to estimate cars promotion using improved ID3 algorithm. International Journal of Advanced Research in Computer and Communication Engineering. 2014 Sep; 3(9):8052–61.
12. Bhattacharya S, Mukhopadhyay D, Giri S. Supply chain management in Indian automotive industry: Complexities, challenges and way ahead. International Journal of Managing Value and Supply Chains. 2014 Jun 1; 5(2):49–62.
13. Hanumanthappa M, Sarakutty TK. Predicting the future of car manufacturing industry using data mining techniques. ACEEE International Journal on Information Technology. 2011; 1(2):27–9.
14. Kruse R, Steinbrecher M, Moewes C. Data mining applications in the automotive industry. 4th International Workshop on Reliable Engineering Computing; National University of Singapore. 2010. p. 23–40.
15. Moskwa JJ, Hedrick JK. Nonlinear algorithms for automotive engine control. IEEE Control Systems Magazine. 1990 Apr; 10(3):88–93.
16. Tayarani-NMH, Yao X, Xu H. Meta-heuristic algorithms in car engine design: A literature survey. IEEE Transactions on Evolutionary Computation. 2015 Oct; 19(5):609–29.
17. CarDeho. Available from: www.cardekho.com
18. Ferrari 458 Speciale A Reviews – CarBuzz. Available from: http://www.carbuzz.com/Ferrari/2015_Ferrari_458-Speciale-A/
19. Ferrai F12tdf review in pictures. Available from: http://www.evo.co.uk/ferrari/f12tdf
20. Ferrari California T Photos and Info – News – Car and river. Available from: http://www.caranddriver.com/ferrari/california-t
21. Aventador LP 750-4 Superveloce Roadster - Lamborghini Nürnberg Aventador LP 750-4 Superveloce Roadster - Lamborghini Nürnberg. Available from: http://www.lamborghini.com/en/models/aventador-lp-750-4-superveloce-roadster/technical-specifications/
22. Komikders.com safety review, security report - SiteSec.co. Available from: http://www.ferrari.com/en_en/?pillarName=ferrari
23. XC60 Specifications Volvo Cars. Available from: http://www.volvocars.com/in/cars/new-models/xc60/specifications
24. Han J, Kamber M, Pei J. Data mining, southeast asia edition: Concepts and techniques. Morgan Kaufman; 2006. p. 740.