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. The buying power of people has increased with the rise in incomes and emerging middle class. The demands for automobiles have increased manifold due to the enhanced infrastructure and ease of access of commercial automobiles for distant markets. Extensive research works have been carried out in this industry to improve features like design, manufacturing, 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. The application of Data Mining can be seen in various fields of Automobile sector which includes manufacturing, production, performance, safety etc. Data Mining techniques such as clustering, fuzzing, 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...
Automobile Engine Performance Analysis using Regression Technique

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.

\[ y = w_0 + w_1x \]

where \( w_0 \) 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.

\[ y = w_0 + w_1x_1 + w_2x_2 \]

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)).

\[ R^2 = 1 - \frac{SCE}{SCT} \left( \frac{n - 1}{n - p - 1} \right) \]

2.2.1 T-TEST

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

The statistical test is given as

\[ t_j = \frac{a_j}{\sigma_j} \]

\( \sigma_j \) is the standard error of the estimated coefficient. The diagonal of the covariance matrix of the estimated coefficients provides its squared value (Equation (3)).

\[ \Omega = \sigma^2(X'X)^{-1} \]

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

\[ \sigma_e = \sqrt{\frac{SCE}{n - p - 1}} \]

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.

\[ F = \frac{SCE/p}{SCE/(n - p - 1)} \]

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

<table>
<thead>
<tr>
<th>Year</th>
<th>Name of the Author(S)</th>
<th>Application</th>
<th>Data Mining Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009</td>
<td>Qian Zhou</td>
<td>Vehicle Report-Stop</td>
<td>Data Mining &amp; Data Warehousing</td>
</tr>
<tr>
<td>2010</td>
<td>Rudolf Kruse, Matthias Steinbrecher and Christian Moewes</td>
<td>Car lifecycle stages</td>
<td>Clustering</td>
</tr>
<tr>
<td>2014</td>
<td>M.Jayakameswaraih and S.Ramakrishna</td>
<td>Car Manufacturing</td>
<td>ID3Algorithm</td>
</tr>
<tr>
<td>2011</td>
<td>Hanumanthappa</td>
<td>No. of cars Manufactured</td>
<td>Linear regression</td>
</tr>
<tr>
<td>2011</td>
<td>S.Gunasekaran and C.Chandraskaran</td>
<td>Automobile industry data</td>
<td>Clustering</td>
</tr>
<tr>
<td>2012</td>
<td>Marco Hulsman, Christoph M. Freidrich and Dirk reith</td>
<td>Sales forecast</td>
<td>Time series analysis and Classical algorithm</td>
</tr>
<tr>
<td>2009</td>
<td>Liu Gaojun and Long Boxue</td>
<td>Sales forecast</td>
<td>Math statistics and neural network</td>
</tr>
</tbody>
</table>
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.

4. Experimental Setup

4.1 Dataset

The dataset of various automobiles was collected from internet sources. 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 1. Methodology of proposed analysis.

Figure 2. Automobile dataset.

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. The selection of principal component is carried out using this Eigen values. Figure 4 shows the Eigen values table of PCA.

![Figure 4. Eigen values table.](image)

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.

![Figure 5. Eigen vector table.](image)

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.](image)

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.](image)
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.

<table>
<thead>
<tr>
<th>Endogenous attribute</th>
<th>Engine Displacement (in cc)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example</td>
<td>320</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.824777</td>
</tr>
<tr>
<td>Adjusted-$R^2$</td>
<td>0.823567</td>
</tr>
<tr>
<td>Sigma error</td>
<td>689.531484</td>
</tr>
<tr>
<td>F-Test (2,317)</td>
<td>746.0375 (0.000000)</td>
</tr>
</tbody>
</table>

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 $\alpha = 0.05$ (or 5%), then the model is considered as significant.

The p-value was found out to be 0.00. Figure 9 shows the 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.
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 (R²) 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 R².

<table>
<thead>
<tr>
<th>Endogenous attribute</th>
<th>Engine Displacement (in cc)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examples</td>
<td>320</td>
</tr>
<tr>
<td>R²</td>
<td>0.839620</td>
</tr>
<tr>
<td>Adjusted-R²</td>
<td>0.838098</td>
</tr>
<tr>
<td>Sigma error</td>
<td>660.713873</td>
</tr>
<tr>
<td>F-Test (3,316)</td>
<td>551.4417 (0.000000)</td>
</tr>
</tbody>
</table>

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](image)

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.

The value of coefficient of determination (R²) 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.

<table>
<thead>
<tr>
<th>Endogenous attribute</th>
<th>Engine Displacement (in cc)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examples</td>
<td>320</td>
</tr>
<tr>
<td>R²</td>
<td>0.922447</td>
</tr>
<tr>
<td>Adjusted-R²</td>
<td>0.923730</td>
</tr>
<tr>
<td>Sigma error</td>
<td>453.486684</td>
</tr>
<tr>
<td>F-Test (3,316)</td>
<td>1288.8225 (0.000000)</td>
</tr>
</tbody>
</table>

Figure 13. Results table coefficient of determination.

The value of coefficient of determination (R²) 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 14 describes the ANOVA table for this case.

![Analysis of variance](image)

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 (R²) 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.

<table>
<thead>
<tr>
<th>Endogenous attribute</th>
<th>Engine Displacement (in cc)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examples</td>
<td>320</td>
</tr>
<tr>
<td>R²</td>
<td>0.827476</td>
</tr>
<tr>
<td>Adjusted-R²</td>
<td>0.825838</td>
</tr>
<tr>
<td>Sigma error</td>
<td>685.272784</td>
</tr>
<tr>
<td>F-Test (3,316)</td>
<td>505.2101 (0.000000)</td>
</tr>
</tbody>
</table>

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](image)

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
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