CA02 Report-Comparison of various Car features to predict Car Prices

Data Mining

B9DA103

Bharat Jethwani 10519364

Table of Contents

1. Introduction

This report would give an accurate idea of what I learned while implementing the project "Comparison of various Car features to predict Car Prices."

2.What did I Learn

For the study we have used two tools, namely:

1. RapidMiner
2. Tableau

I have tried to explain what I learned through this project; the whole section divided into six parts with each part containing two sub

- What I understood about each phase of crisp-dm.
- What concepts and techniques I learned during this project.

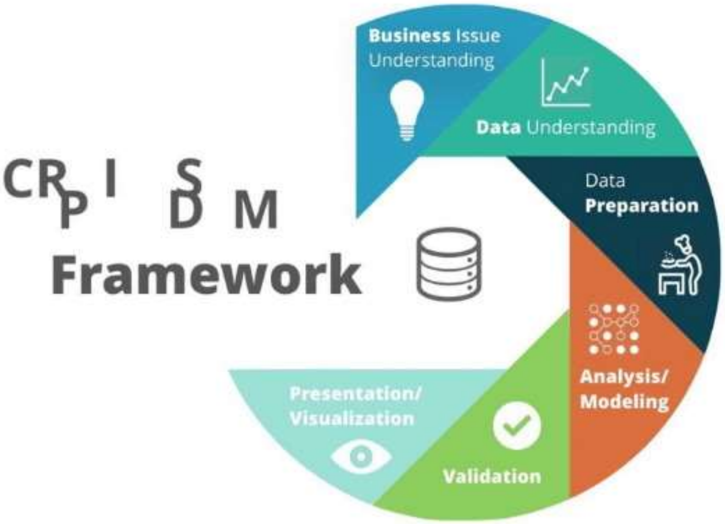


Fig.1 Crisp-DM

2.1 Business Understanding

2.1.1 What I understood about this phase:

- This is the most crucial phase of any data mining project; this phase is all about developing a preliminary plan designed to achieve the project objective.
- It involves analyzing the background, i.e., understanding the goals of the project and discovering important and hidden factors.
- Stating the primary objectives and ensuring the project answers to these objectives and not the wrong questions.
- Defining success criteria and making sure the project is goal specific while not setting unachievable goals.
- List the assumptions and risks involved in the project.

2.1.2 Concepts and techniques I learned during this phase

1. How we used the idea of Bussiness understanding for our project:

I have always had a craze about cars since childhood, and with my interest growing in learning about different regression techniques, we decided to explore around Cars.

Most of the families before buying a car always wonder, is it worth enough to pay x amount for a car? , do the features promised, affect a car's price to that extent? Our study revolves around these two questions. Our primary objective is,

What features most affect a car's worth, and to what extent?

Other sub-questions included:

- Can horsepower affect the car price?
- Can the price of the car be predicted based on its features?
- Does the width and height of a car affect the price prediction?

2.Regression:

- For comparing factors affecting a car price and comparing car prices, we used regression modeling techniques.
- It is a type of Supervised ML model. In this type of learning, machine(model) is trained on well-labeled data, i.e., some data is already tagged with the correct answer. In real life, it could be related to an example of "Student training under the supervision of a teacher."
- Regression when the target variable is continuous, like weight or height.

2.2 Data Understanding

2.2.1 What I understood about this phase:

- This phase involves getting familiar with the data, discovering insights, detecting different useful information from subsets to create an interesting hypothesis.
- Collecting and loading the necessary data, describing the data by stating the number of records and fields.
- Explore the data by comparing attributes with each other by forming visualizations and reports.
- One of the critical initial steps is to check the data quality and missing values.

2.2.2 Concepts and techniques I learned during this phase

1.what techniques we used for this project, and how?

For this project, we used the automobile dataset, which we extracted from "https://archive.ics.uci.edu/ml/datasets/automobile" and this dataset explores few cars from every range and top makers, be it a hatchback, sedan or SUV and their features. Dataset has 26 attributes and

205 rows with 59 missing values. We started with reading the data using read CSV operator from rapid miner as shown in the below figure.

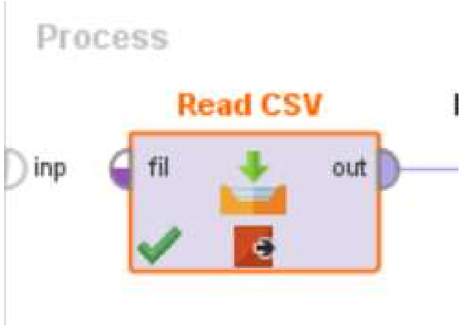
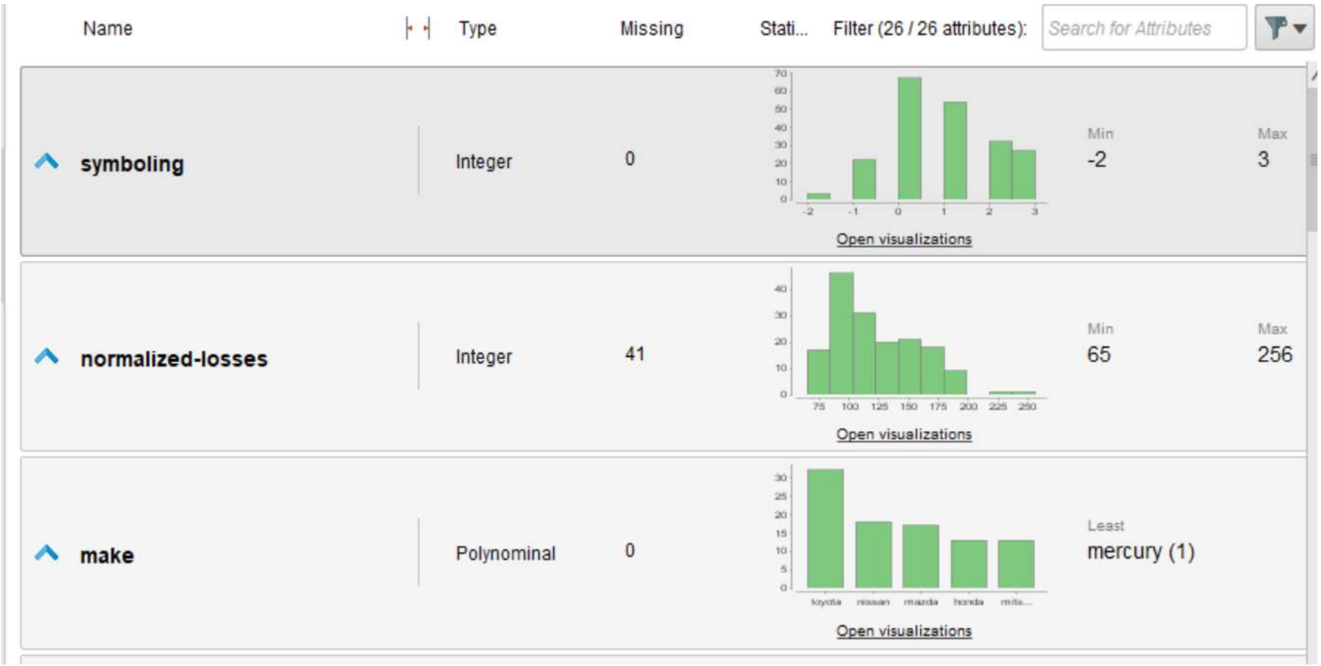


Fig2. Reading automobile dataset

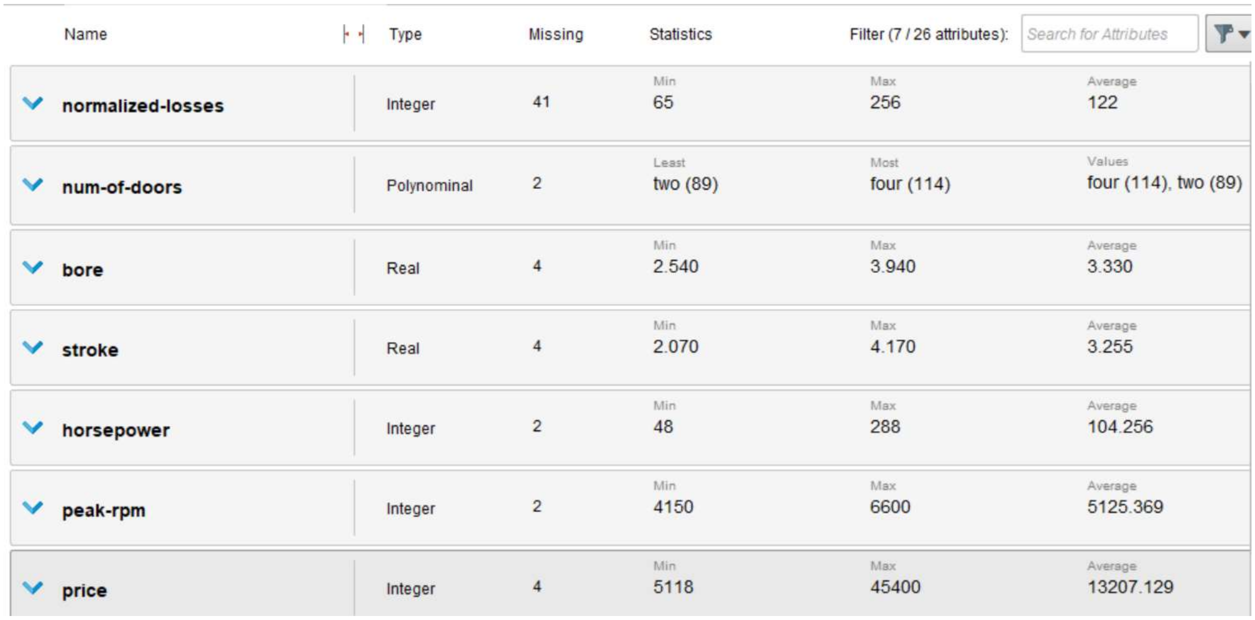
This operator gives us a detailed idea about the dataset by describing data type and statistics for each attribute. Along with that it provides quick visualization explaining distribution across the dataset for continuous variables and data splitting along each group for categorical variables.



Filter (26 / 26 attributes): Search for Attributes

Name	Type	Missing	Stati...		
symboling	Integer	0	Open visualizations	Min -2	Max 3
normalized-losses	Integer	41	Open visualizations	Min 65	Max 256
make	Polynominal	0	Open visualizations	Least mercury (1)	

Fig3. Understanding data using statistics tab or RapidMiner.



Name	Type	Missing	Statistics		
normalized-losses	Integer	41	Min 65	Max 256	Average 122
num-of-doors	Polynominal	2	Least two (89)	Most four (114)	Values four (114), two (89)
bore	Real	4	Min 2.540	Max 3.940	Average 3.330
stroke	Real	4	Min 2.070	Max 4.170	Average 3.255
horsepower	Integer	2	Min 48	Max 288	Average 104.256
peak-rpm	Integer	2	Min 4150	Max 6600	Average 5125.369
price	Integer	4	Min 5118	Max 45400	Average 13207.129

Fig4. Attributes with missing values

2.3 Data Preparation

2.3.1 What I understood about this phase:

- Includes all the methods and activities to construct the final dataset that would be fed to the model.
- It starts with selecting necessary and essential data using manual or automated feature selection techniques to avoid high bias and high variance.
- Cleaning data, i.e., dealing with missing values by either deleting them or replacing them with necessary techniques and detecting outliers and dealing with them.
- Create derived attributes if necessary and combining appropriate subsets.
- Formatting the data according to modeling input requirements could be achieved using normalization, discretization, and sampling.

2.3.2 Concepts and techniques I learned during this phase

1.what techniques we used for this project, and how?

- With the Knowledge learned about the data, we had decided to deal with missing values and select necessary attributes with the hope of achieving a bias-variance trade-off.
- As mentioned in the above phase, we had 59 missing values in total. We decided to replace missing values using the "impute missing values" operator in RapidMiner. For the same, we used KNN to predict the missing values.
- Once, dealing with the missing values, we normalized the data and used the auto-model feature to unfold the suspense of which model would be the best for predicting prices. And Random forest emerged as the victorious model.

- Once we found out the best model for the Automobile dataset, we had to extract the best features for our purpose. For doing the same, we used an optimized selector operator with a cross-validation technique with Random forest as a model. As a result of that six attributes were extracted.

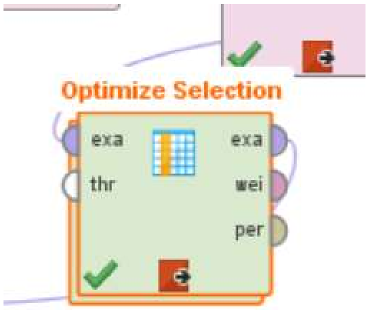


Fig5. Optimize selector for feature extraction

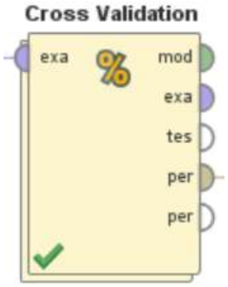


Fig6. Cross-Validation technique

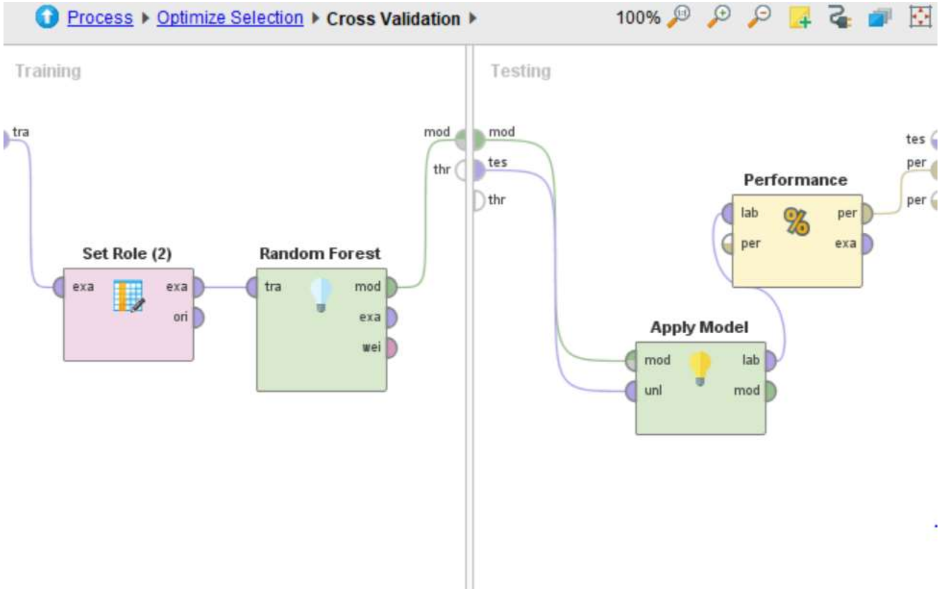
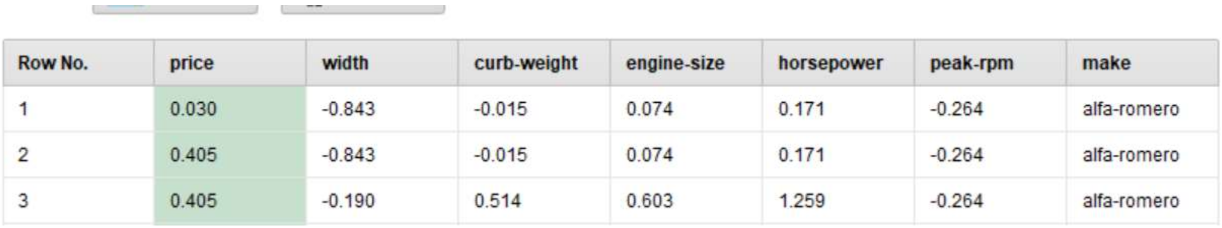


Fig7. Random forest as optimize selector



Row No.	price	width	curb-weight	engine-size	horsepower	peak-rpm	make
1	0.030	-0.843	-0.015	0.074	0.171	-0.264	alfa-romero
2	0.405	-0.843	-0.015	0.074	0.171	-0.264	alfa-romero
3	0.405	-0.190	0.514	0.603	1.259	-0.264	alfa-romero

Fig8. Attributes selected after feature selection.

2. Feature Selection

I believe machine learning works on a simple principle; if you give it garbage, the output will stink; in this metaphor, garbage refers to noise (non-relevant data). Imagine a large dataset with such data and some useful data. It becomes very important to extract the required important data for the model to learn better.

- It allows the model to train faster.
- Improves accuracy
- Reduce the complexity of the model.
- Avoids overfitting and high bias.
- Gaining a better understanding of the features and their relationship.

3. Cross-Validation:

- It is one of the methods to estimate and improve model performance by comparing train and test accuracy. It also helps in selecting a suitable model for a certain dataset by comparing test errors for each model.
- To estimate the performance, it splits the data set into k-parts with each part being called as a fold. Evaluation is done by taking each fold as test set and the model is trained on remaining folds, thus testing the whole dataset.
- Each run gives us a score, thus ending up with k scores. Evaluation is done by taking mean and standard deviation for k scores.
- The value of k should be taken very carefully as the whole model would run k times increasing the computation time and wasting resources.
- Cross-validation helps in analyzing model performance in the following ways
  1. If test accuracy is greater (around 10%) than training accuracy, then its training problem i.e., the model, didn't learn well and made unnecessary assumptions resulting in high bias and a classic case of underfitting. The solution for such cases is using parametric models, feature selection.

If training accuracy is greater (around 10%) than testing accuracy, then there is a problem of high bias, i.e., confusion caused because of too much learning a classic case.

4. Why KNN for imputing missing values:

- KNN is a model that works on matching a point with its closet neighbors in multi-dimensional space.
- It can be used for all kinds of data types, continuous, categorical, etc. since our data had different data types, this was beneficial.
- Also, it's easy to compute and learn.
- The main reason to use KNN for handling missing values was it approximates the value to the points that are closest.

5. Bias-Variance trade off

- The objective of any supervised learning model is to obtain low bias and low variance. Parametric (linear) machine learning usually tend to have a high bias, but low variance while non-parametric (non-linear) tend to have low bias and high variance.
- There can be no escape from the relationship of bias and variance i.e., increasing bias would decrease variance and vice versa.
- Examples for achieving trade-off:
  1. KNN has low bias and high variance, but the bias can be increased by increasing the number of k i.e., increasing the number of neighbors.
  2. SVM has low bias and high variance, but increasing the c parameter of margin in training data would increase the bias.

2.4 Data Modeling

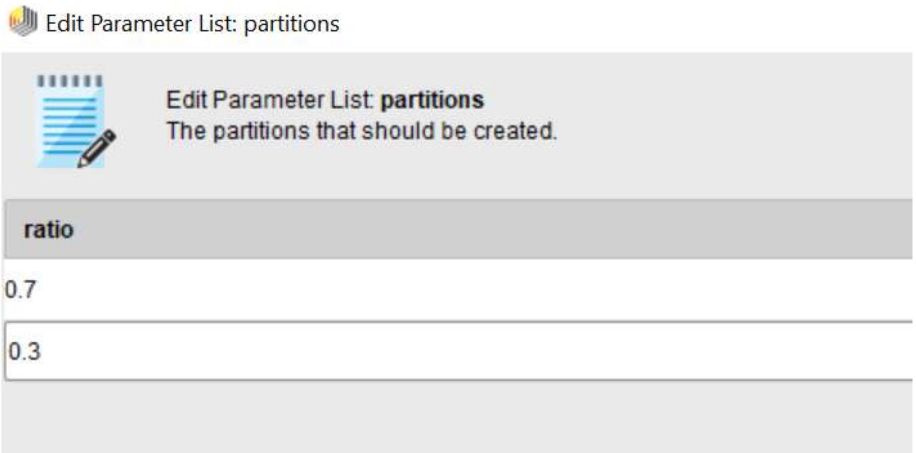
2.4.1 What I understood about this phase:

- This phase includes selecting necessary models or a combination of models initialing estimating the performance of various models using techniques like cross-validation.
- Split train and test data in the required ratio using multiple methods.
- Machine learning analyst would then fir the model.
- Asses the model by the success criteria initially stated.

2.4.2 Concepts and techniques I learned during this phase

1.what techniques we used for this project, and how?

- After cross-validation we had to split data into test and train, for the same we used split data operator in RapidMiner, splinting our dataset into 70% trainset and 30% testset.
- It was clear that Random forest/ Gradient tree bosted would the best choice for our project, but we had confidence with the extracted features KNN would work best.
- After comparison and trying various algorithms we decided to go with KNN as it gave the best results.



Edit Parameter List: partitions

Edit Parameter List: **partitions**
The partitions that should be created.

ratio
0.7
0.3

Fig9. Splitting data

2. Random Forest:

- It is a model which is used for both Regression and classification.
- For the same, it uses multiple decision trees along with bootstrap, Aggregation, which is commonly known as bagging.

3. The idea behind using KNN:

- It is a simple example used for prediction of a numerical attribute using the nearest neighbor data points.
- It is one of the best algorithms used for pattern recognization and Predicting numerical targets.

2.5 Evaluation

2.5.1 What I understood about this phase:

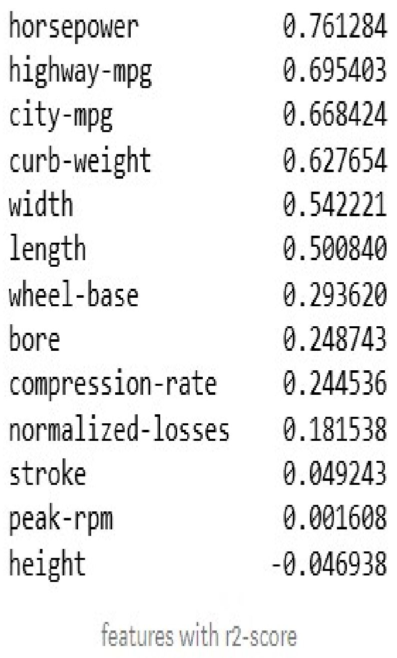
- The necessary evaluation technique should be selected, for example, evaluation-classification for classification problems and evaluation-regression for regression problems.
- Machine learning analyst should validate If objectives are attained with desired success criteria or not.
- Also, attaining Bias-Variance Tradeoff to avoid overfitting and underfitting so that model could generalize better with new unknown data using cross-validation techniques.
- If NO, making the necessary changes by repeating the necessary phases.
- If YES, summarizing evaluation results using reports and visualizations.
- Moving to deployment if a recurring project.

2.5.2 Concepts and techniques I learned during this phase

1.what techniques we used for this project, and how?

- As our project is based on regression learning, we used the evaluation regression operator in data mining.

- We mainly concentrated on RMSE and squared correlation with values being 0.252 and 0.955, respectively.
- The result we got was much better than we expected. With horsepower, engine weight, and curb weight being the top three features affecting car price.



```
horsepower           0.761284
highway-mpg          0.695403
city-mpg             0.668424
curb-weight          0.627654
width                0.542221
length               0.500840
wheel-base           0.293620
bore                 0.248743
compression-rate     0.244536
normalized-losses    0.181538
stroke               0.049243
peak-rpm             0.001608
height              -0.046938
```

features with r2-score

Fig10. r2 score of various features if tested individually

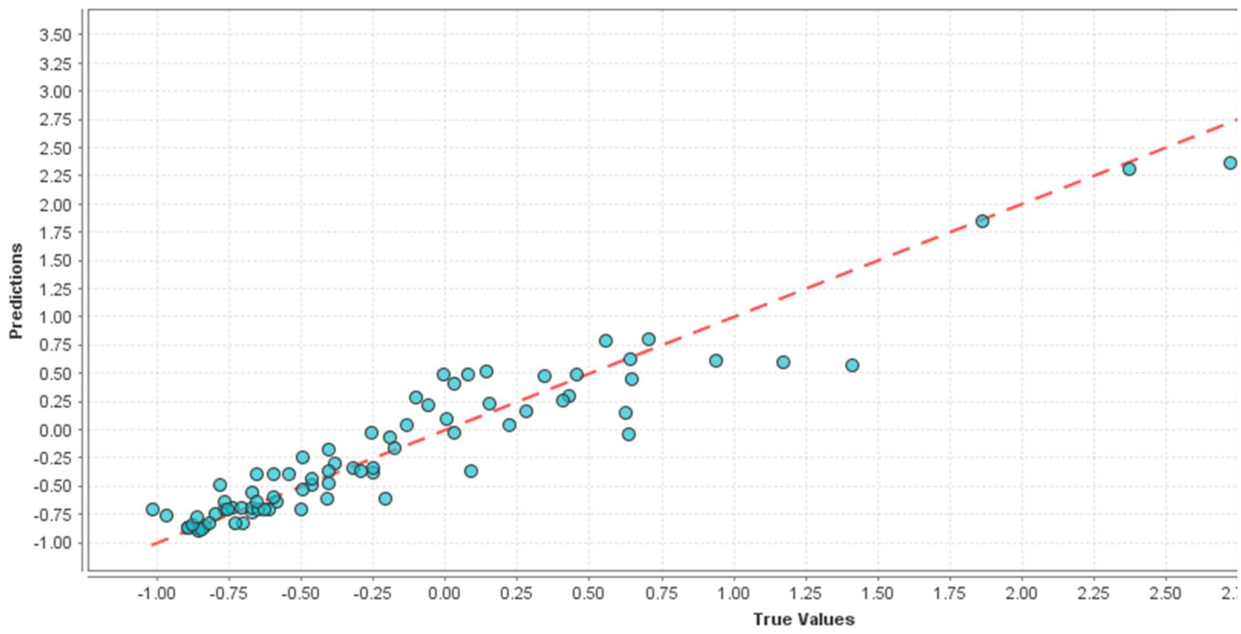
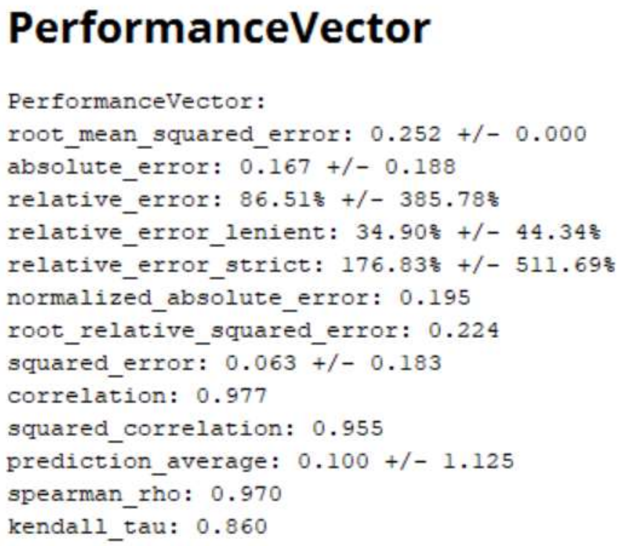


Fig11. Actual VS Predicted Price



PerformanceVector

```
PerformanceVector:
root_mean_squared_error: 0.252 +/- 0.000
absolute_error: 0.167 +/- 0.188
relative_error: 86.51% +/- 385.78%
relative_error_lenient: 34.90% +/- 44.34%
relative_error_strict: 176.83% +/- 511.69%
normalized_absolute_error: 0.195
root_relative_squared_error: 0.224
squared_error: 0.063 +/- 0.183
correlation: 0.977
squared_correlation: 0.955
prediction_average: 0.100 +/- 1.125
spearman_rho: 0.970
kendall_tau: 0.860
```

Fig12. Performance Vector

2.Evaluation metrics for Regression:

Mean Squared Error (MSE): One of the most commonly used evaluation metrics for regression problems. It is the average of the squared difference between the target value and the predicted value. It penalizes every small error indicating how bad the model is performing.

Root Mean Squared Error (RMSE): One of the most widely used evaluation technique for regression Determination challenges, it is the square root of averaged squared difference between the target value and value predicted. It penalizes large errors and is mostly used when there are chances for large errors.

Mean Absolute Error (MAE): It is the absolute difference between the target value and the predicted value. It is robust towards outliers but does not penalize errors as strictly as MSE. Not suitable for projects where outliers need to be addressed.

R2 Error: It is also known as the coefficient of determination, which compares the current model with a constant baseline indicating how the model is performing. A constant baseline is the mean of data by drawing the line at the mean. R2 is always less than or equal to 1.
Adjusted R2: It is an advanced version of R2, It is not improving the model but improving model evaluation. It was always less than r2 and showed improvement when there is a real improvement.

2.6 Deployment

2.6.1 What I understood about this phase:

- Develop a strategy for deployment using phase 1, phase 2, and phase 5.
- If the model runs on a frequent basis developing monitoring and management plans using various tools or scripts.
- Produce a detailed final report of the findings and review the whole project by writing the abstract and summary along with conclusions.

2.6.2 Concepts and techniques I learned during this phase

1.what techniques we used for this project, and how?

- We plan to deploy this project as a GUI or web APP in future, for the same we would be using Ansible and Python.

3. Conclusion:

Not only this project, but the whole module helped me to understand CRISP-DM methodologies and various techniques used in each phase. I plan to use this Knowledge for various projects in the future. As far as this project is concerned, I would like to conclude with:

- Makers do affect the car price
- Horsepower is one of the critical features affecting the car price
- The width of a car plays an important role in car pricing.