### CAR PERFORMANCE PREDICTION

Ajith Kumar, Mithilesh Kumar Yadav H, Kavi Bharathi

Sona College of Technology,Salem,Tamin Nadu,Inda

**ABSTRACT:**

A car performance prediction has been a highinterest research area, as it requires noticeable effort and knowledge of the field expert. Considerable number of distinct attributes are examined for the reliable and accurate prediction. To build a model for predicting the price of used cars in Bosnia and Herzegovina, we applied three machine learning techniques (decision tree). The data used for the prediction was collected from the kaggle.Respective performances of different algorithms were then compared to find one that best suits the available data set. The final prediction model was integrated into Web application. Furthermore, the model was evaluated using test data and the accuracy of 86 was obtained.

Keywords – car performance prediction, support vector machines, classification, machine learning.

**1 INTRODUCTION**

**1.1 OVERVIEW**

Predicting the performance level of cars is an important and interesting problem. The main goal of the current study is to predict the performance of the car to improve the certain behavior of the vehicle. This can significantly help to improve the systems fuel consumption and increase the efficiency. The performance analysis of the car based on the engine type, no of engine cylinders, fuel type and horsepower etc. These are the factors on which the health of the car can be predicted. It is an on-going process of obtaining, researching, analyzing and recording the health based on the above three factors. The performance objectives like mileage, dependability, flexibility and cost can be grouped together to play a vital role in prediction engine and engine management system. This approach is the very important step towards understanding the vehicles performance.

**1.2 PURPOSE**

It is an important to analyse the factors using number of well-known approaches of machine learning algorithms like, decision tree and to improve the vehicle performance efficiency. The range, durability and longevity of automotive traction batteries are ‘hot topics’ in automotive engineering. And here we consider a performance in mileage. To solve this problem, we will develop the models, using the different algorithms and neural networks. We will then see decision tree algorithm predicts car performance(Mileage) with higher accuracy.

**2 LITERATURE SURVEY**

**2.1 Existing problem**

Computer simulation tools can give early indicators of key vehicle characteristics. In traditional hybrid vehicles, this is important in designing for optimal fuel consumption; in plug-in hybrids and pure electric vehicles, it is critical for accurate prediction of range, a key market qualifier. There are a variety of techniques, typically operating at different levels of fidelity and employing different modelling philosophies. This paper develops on earlier work, exploring conventional and `backward' techniques in the context of current NEDC-based UNECE vehicle testing standards and the proposed replacements based on the World Light Test Procedure. Model sensitivities for A, C and D-segment vehicles are quantified and this is used to explore aspects where accurate models are key and where lower-fidelity representative models are appropriate. The paper also explores the sensitivity of predictions to `PID control' driver models, and discusses the effect of cycle-following tolerance on predictions. Finally, the paper proposes new standards - suitable for simulation or real-world testing - for a common quantification of in-use battery lifetime. The use of these techniques and the sensitivity analysis methods on a representative simulation model is demonstrated as a case study, and the impacts on battery management strategy design are explored.

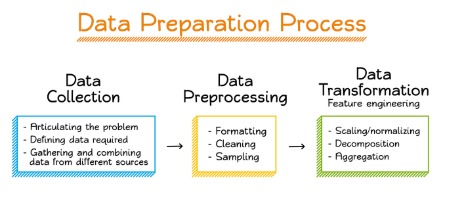
**2.2 Proposed solution**

In this paper, we investigate the application of supervised machine learning techniques to predict the price of used cars in Mauritius. The predictions are based on historical data collected from daily newspapers. Different techniques like decision tree analysis have been used to make the predictions.

The predictions are then evaluated and compared in order to find those which provide the best performances. A seemingly easy problem turned out to be indeed very difficult to resolve with high accuracy. This method provided comparable performance. In the future, we intend to use more sophisticated algorithm.

**3 THEORITICAL ANALYSIS**

**3.1 Block diagram**



**3.2 Hardware / software designing**

A. Hardware requirements The following hardware was used for the implementationof the system:

• 4 GB RAM • 10GB HDD • Intel 1.66 GHz Processor Pentium 4

B. Software requirements The following software was used for the implementationof the system:

• Windows 8/9/10 • Python

**4 EXPERIMENTAL INVESTIGATIONS**

**Data Collection**

we collected dataset from kaggle dataset named as car performance .csv

**Data Processing**

**Import the libraries**

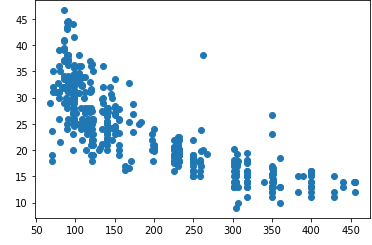
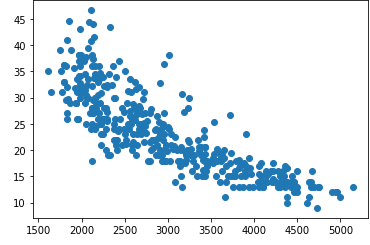
we are importing necessary libraries for our dataset

**Import the dataset**

we are importing dataset with help of pandas libraries

**Data Visualization**

This is the first and foremost step where they will get a high level statistical overview on how the data is. and some of its attributes like the underlying distribution, presence of outliers, and several more useful features.

**Taking Care Of Missing Data**

Sometimes you may find some data are missing in the dataset. We need to be

equipped to handle the problem when we come across them. Obviously youcould remove the entire line of data but what if you are unknowingly removingcrucial information? Of course we would not want to do that. One of the mostcommon ideas to handle the problem is to take a mean of all the values of the

same column and have it to replace the missing data.We will be suing dataset.isnull().any() method to see which column hasmissing values.Word “True” that the particular column has missing values ie, in our datasetage and salary column has missing valuesWe can replace the missing values by, mean median or mode by using fillnamethod.

**Label Encoding**

Sometimes in the dataset we will find textual data like names, countries

states, then the machine cannot do mathematical operations or cannotunderstand the textual data. So the textual data are to be converted in tonumerical format which is called as label encoding. we make use of labelEncoder class to convert textual data in to Numerical data. In the givendataset has textual data so we will be converting that particularcolumns textual data to numerical values.

**One Hot Encoding**

To accomplish the task, we will import yet another library called

OneHotEncoder.Next we will create an object of that class, as usual, and assign it to

onehotencoder. OneHotEncoder takes an important parameter called

categorical\_features which takes the value of the index of the column of

categories.we will use fit\_transform OneHotEncoder and additionally include toarray()

If you check your dataset now, all your categories will have been encoded to

0s and 1s.

**Feature Scaling**

The final step of data preprocessing is to apply the very important feature

scaling. It is a method used to standardize the range of independent variables

or features of data.A lot of machine learning models are based on Euclidean distance. If, for

example, the values in one column (x) is much higher than the value in another

column (y), (x2-x1) squared will give a far greater value than (y2-y1) squared.

So clearly, one square difference dominates over the other square difference.

In the machine learning equations, the square difference with the lower value

in comparison to the far greater value will almost be treated as if it does not

exist. We do not want that to happen. That is why it is necessary to transform

all our variables into the same scale. There are several ways of scaling the

data. One way is called Standardization which may be used. For every

observation of the selected column, our program will apply the formula of

standardization and fit it to a scale.

**Splitting Data Into Train And Test**

Now we need to split our dataset into two sets — a Training set and a Test set.

A general rule of the thumb is to allocate 80% of the dataset to training set

and the remaining 20% to test set. For this task, we will import test\_train\_split

from model\_selection library of scikit.Now to build our training and test sets, we will create 4 sets— X\_train(training part of the matrix of features), X\_test (test part of the matrix of

features), Y\_train (training part of the dependent variables associated with the

X train sets, and therefore also the same indices) , Y\_test (test part of the

dependent variables associated with the X test sets, and therefore also the

same indices). We will assign to them the test\_train\_split, which takes the

parameters — arrays (X and Y), test\_size (if we give it the value 0.5, meaning

50%, it would split the dataset into half. Since an ideal choice is to allocate

20% of the dataset to test set, it is usually assigned as 0.2. 0.25 would mean

25%, just saying)

**Model Building**

**Training And Testing The Model**

This is uses any one of the models that we had chosen in step 3. Once themodel is trained we can use the same trained model to predict using the testing data i.e. the unseen data.

**Evaluation**

Once this is done we can calculate the performance of the linear regression model by calculating accuracy.The above statement will calculate the r2 value of Y\_test and y\_ pred. If this valueis above 80 then it means that the model got trained well

If you are solving classification problem. You can use any classification algorithm

which is mentioned in step 3 for example you can use Decision tree as shown below.

Once this is done we can develop a confusion matrix, this tells us how well our

model is trained. A confusion matrix has 4 parameters, which are ‘True positives’,

‘True Negatives’, ‘False Positives’ and ‘False Negative’. We prefer that we get

more values in the True negatives and true positives to get a more accurate model.

The size of the Confusion matrix completely depends upon the number of classes.

· True positives: These are cases in which we predicted TRUE and our

predicted output is correct.

**Application Building**

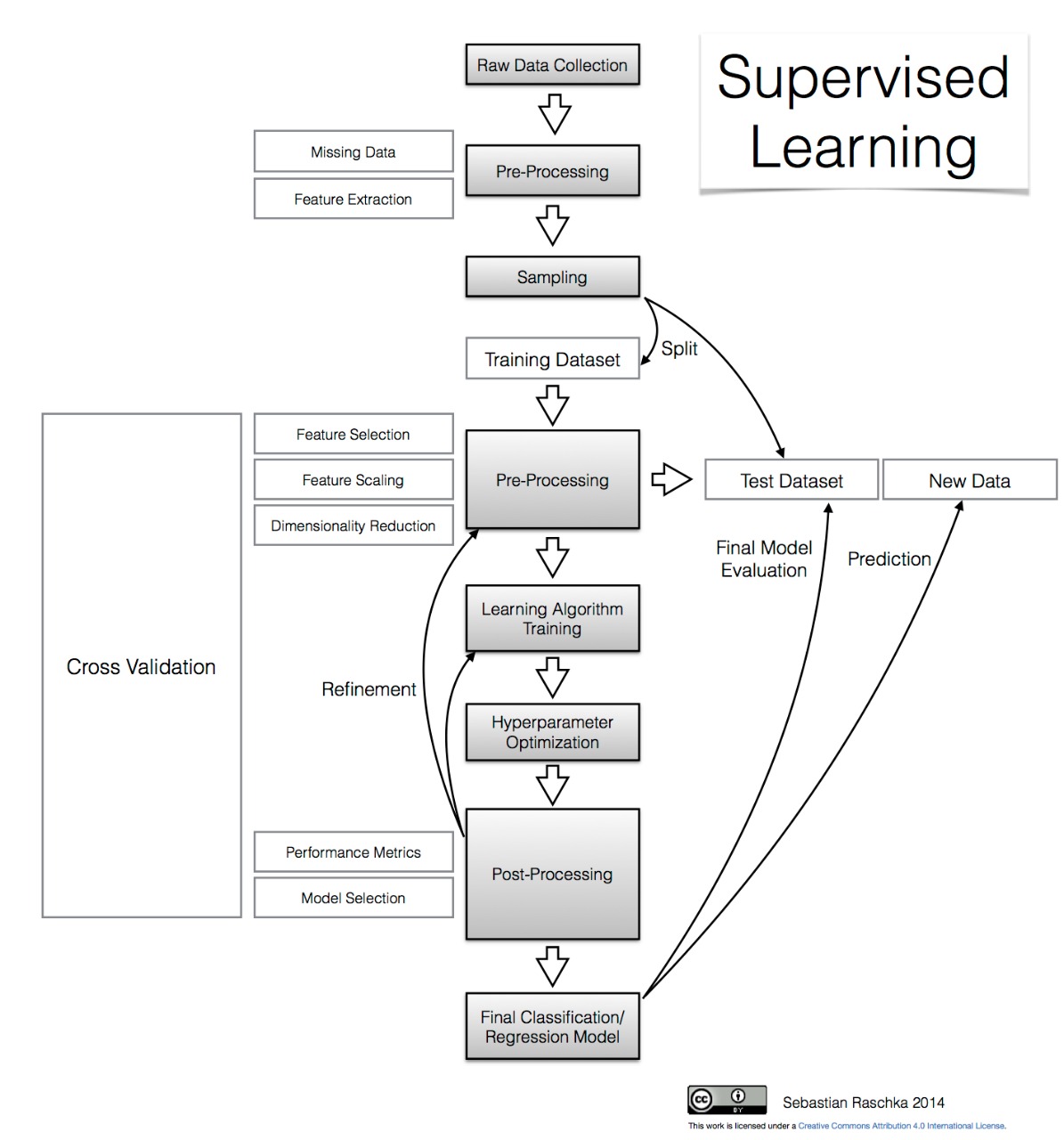
**Create An HTML File**

This is the basic HTML page for our Project. H1 tag is used to give heading to the project. As I have mentioned there are 8 input or 8 independent variables we have

created 8 text input fields in the html page. A button is used to send these values to the model files this functionality will be written in the python file app.py. the model predicts the value and is displayed on the{{ prediction\_text }}filed.

**Build python code**

We will be using python for server side scripting. Let’s see step by step process for writing backend code.

**5 FLOWCHART**

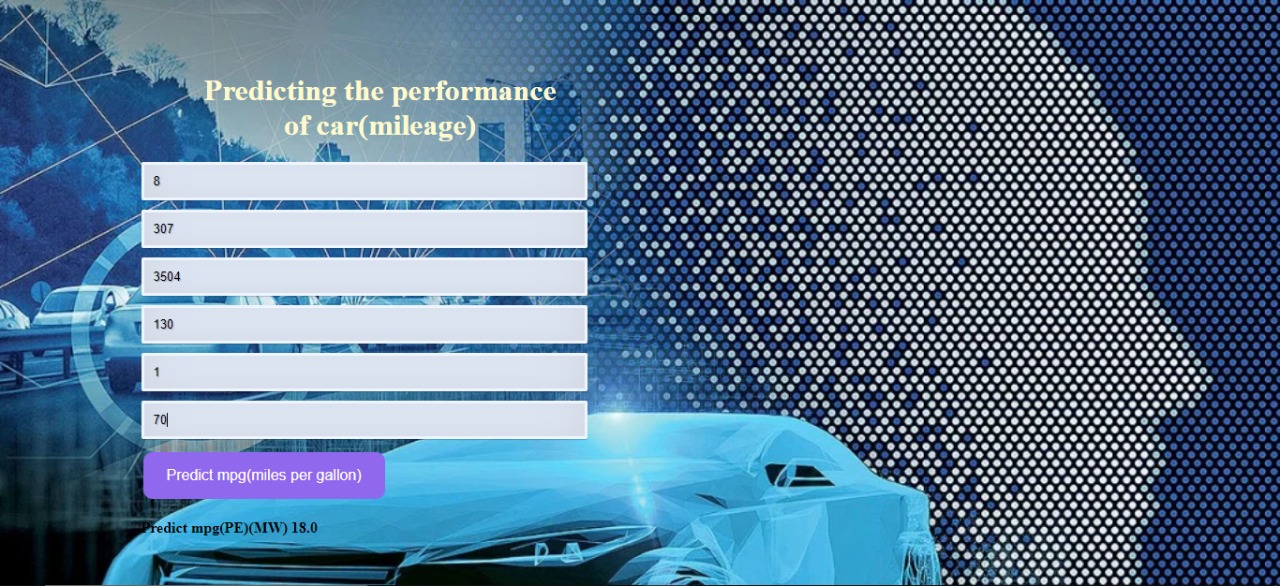
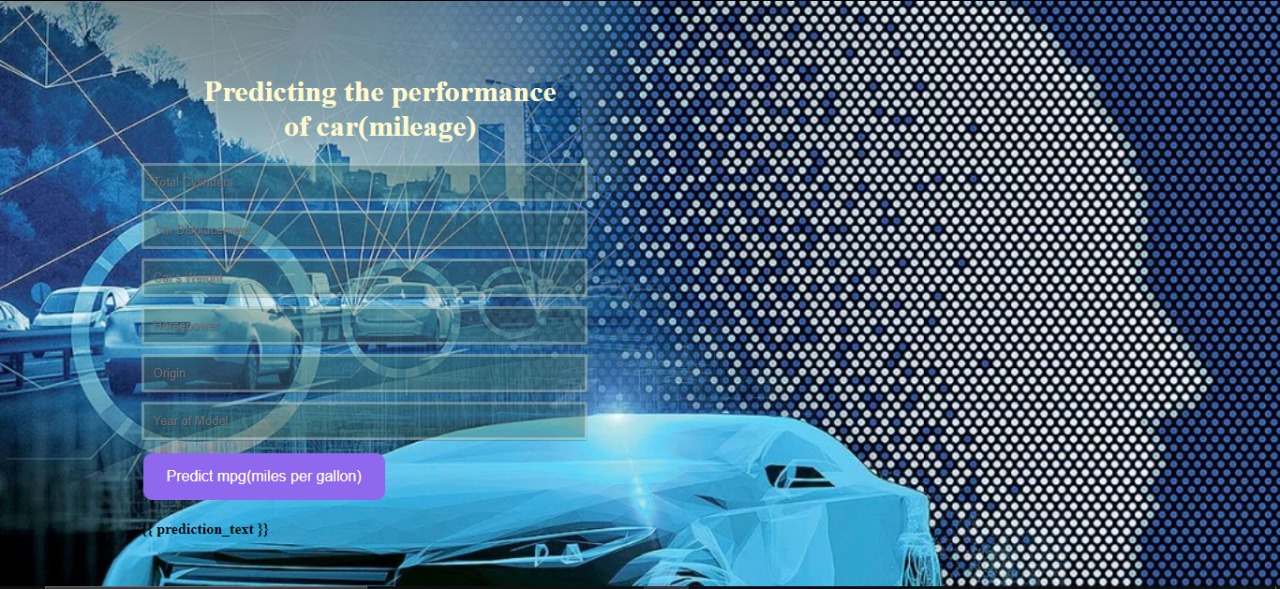
**6 RESULT**

we have got the desired result of more than 80% accuray for prediction of car mileage by using decision tree algorithm.

**7 ADVANTAGES & DISADVANTAGES**

with this we can predict mileage in effecient way some times it may predict wrongly.

**8 APPLICATION**



**9 CONCLUSION**

The high number of attributes that should beconsidered for the accurate prediction. The major stepin the prediction process is collection andpreprocessing of the data. In this research, standardize and cleandata to avoid unnecessary noise for machine learning algorithms. Data cleaning is one of the processes that increases prediction performance, yet insufficient for the cases of complex data sets as the one in this research. Applying random forest and other machine algorithm on the data set accuracy was less than 80%. Therefore, Decision tree machine learning algorithms has been proposed and this combination of MLmethods gains accuracy of 86%. This is significant improvement compared to other machinelearning method approach. However, the drawbackof the proposed system is that it consumes much more computational resources than other machinelearning algorithm used cars data sets and validate the proposed approach

10  **FUTURE SCOPE**

In the future, this accuracy of these models could be improved through further feature engineering. For example, I could create an additional column that used the horsepower and weight data to create a power-to-weight ratio feature. Additionally, I could extend the grid search parameter space to include more of the hyper parameters included in the default MLP Regressor implementation. Finally, I would be excited to see how my predictions would fair if I used a more powerful library, such as Tensor Flow, as the machine learning backend.

11 **BIBLOGRAPHY**

[1] NATIONAL TRANSPORT AUTHORITY. 2014. Available from:

<http://nta.gov.mu/English/Statistics/Pages/Archive.aspx> [Accessed 15 January

2014].

[2] MOTORS MEGA. 2014. Available from:

<http://motors.mega.mu/news/2013/12/17/auto-market-8-decrease-sales-newcars/> [Accessed 17 January 2014].

[3] LISTIANI, M., 2009. Support Vector Regression Analysis for Price Prediction

in a Car Leasing Application. Thesis (MSc). Hamburg University of

Technology.

[4] RICHARDSON, M., 2009. Determinants of Used Car Resale Value. Thesis

(BSc). The Colorado College.

[5] WU, J. D., HSU, C. C. AND CHEN, H. C., 2009. An expert system of price

forecasting for used cars using adaptive neuro-fuzzy inference. Expert Systems

with Applications. Vol. 36, Issue 4, pp. 7809-7817.

[6] DU, J., XIE, L. AND SCHROEDER S., 2009. Practice Prize Paper - PIN

Optimal Distribution of Auction Vehicles System: Applying Price Forecasting,

Elasticity Estimation and Genetic Algorithms to Used-Vehicle Distribution.

Marketing Science, Vol. 28, Issue 4, pp. 637-644.

[7] GONGGI, S., 2011. New model for residual value prediction of used cars based

on BP neural network and non-linear curve fit. In: Proceedings of the 3

rd IEEE

International Conference on Measuring Technology and Mechatronics

Automation (ICMTMA), Vol 2. pp. 682-685, IEEE Computer Society,

Washington DC, USA.