American Journal of Electronics & Communication

Krishnendu Ghosh et. al., American Journal of Electronics & Communication, Vol. III (3), 1-5

Application of Machine Learning Classifier in Automobile Domain

Krishnendu Ghosh
Department of CSE, UEM Kolkata
krg8317@gmail.com

Rajendrani Mukherjee
Department of CSE, UEM Kolkata
rani.mukherj@gmail.com

Abstract - Over the last decade, machine learning has seen widespread application in several domains. Automobile or transportation industry has leveraged different predictive analytics techniques to cater customer requirements, user satisfaction, manufacturer profit analysis etc. In this paper, KNN algorithm is applied on a standard dataset which has been taken from UCI machine learning repository. The automobile dataset contains information regarding car mileage, engine, weight, price etc. As kNN classifier was applied on this benchmark dataset, accuracy was determined from confusion matrix to gain valuable insights.

Keywords – automobile, classifier, data, accuracy, requirements

1. Introduction

Intelligent data analysis (IDA) is playing a pivotal role coupled with machine learning to make prominent decisions/predictions. Since its inception, ML is widely applied in healthcare domain, financial domain, social media, agriculture, transportation sector etc. This research explores the application of AI in automobile domain using a standard benchmark dataset from UCI machine learning repository.

The dataset has information about car engine, fuel type, number of cylinders, price etc. The dataset was cleaned and pre-processed before applying ML classifier. The price is predicted based on independent variables. As RMSE (Root Mean Square Error) values were tabulated, following features appeared as major contributor - 'highway-mpg', 'curb-weight', 'horsepower', 'width' and 'city-mpg'. Different k values were tried as KNN classifier was explored. This kind of analysis will help to understand what are the important factors regarding an automobile before taking a final decision for profit.

The organization of the paper is as follows- Section 2 surveys the literature while Section 3 details the implementation steps. The implementation section contains code snippets and graphical representation for better visualization. Section 4 concludes the work with future scope.

2. Literature Survey

Machine Learning has seen fruitful application in many businesses. Numerous IDA (Intelligent Data Analysis) techniques have enriched the domain of AI [3] [4] [5] and helped in effective decision making.

John T. Behrens pointed the differences between classical data analysis and exploratory data analysis [7]. ETL (Extract-Transform-Load) technology was explored since 2004 [2]. Ehrlinger et al. investigated several data quality management tools [1]. In 2018, V Elliott described coding process for qualitative data analysis [6].

The history of application of ML techniques in automobile industry dates long back [8]. Yaqi et al. utilized random forest and nearest

neighbour classifier for detecting insurance frauds [9].

3. Implementation

In this section, detail implementation is described step by step.

The dataset was split into training and testing data using the function “train_test_split”. We are using 85% of the data for training and rest 15% for validation. The target data ‘price’ is put in a separate frame y. The price is predicted depending on other independent variables.

Following code snippets were run to perform these operations -

```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x_data, y_data, test_size=0.15,
                                                    random_state=1)
print("number of test samples :", x_test.shape[0])
print("number of training samples:", x_train.shape[0])
```

```
number of test samples : 31
number of training samples: 174
```

```
y_data = car['price']
x_data=car.drop('price', axis=1)
```

In the next step, the model is trained using 'horsepower', 'curb-weight', 'engine-size' and 'width’ as features. Price was predicted using train data. The distribution of the predicted values of the training data appears in the following figure Fig. 1.

The x axis represents price in dollars. From the figure it is evident that the model seems to be doing well in learning from the training dataset.

```
lr = LinearRegression()
lr.fit(x_train[['horsepower', 'curb-weight', 'engine-size', 'width']], y_train)
```

```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=False)
```

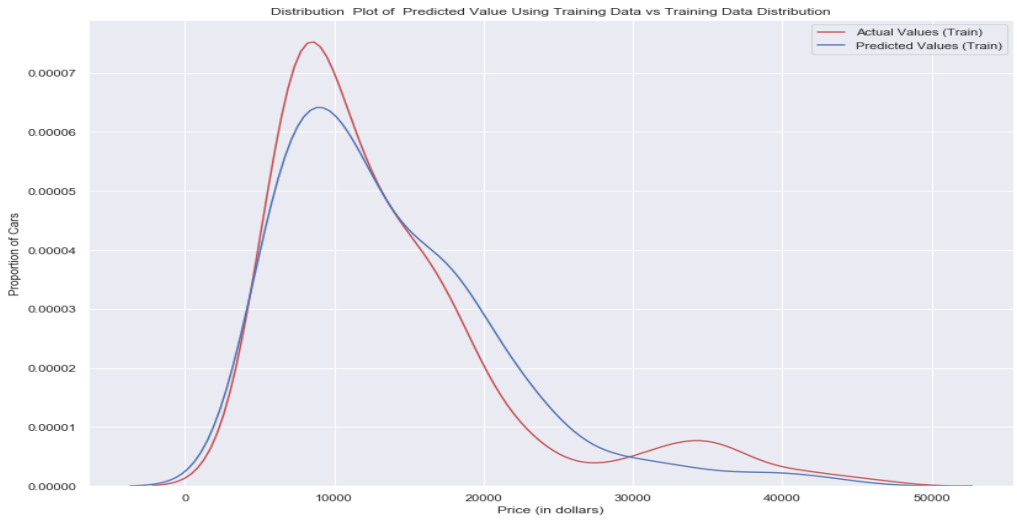

Fig.1 Prediction of car price using train data

The numeric values are normalized to prevent outliers when measuring squared errors. Next, KNN classifier is applied on the dataset taking the training column, target column, the data frame object, and a parameter for the k value. The k value is chosen as 3, 5, 7 and 9.

All k values were applied and the RMSE (Root Mean Square Error) results were plotted. Fig. 2 represents the plots. RMSE values range from 4,000 dollars up to 11,000 dollars. The top 5 features based on the average of all RMSE values for each *k* value are chosen. From the table it is evident that the top 5 features are - 'highway-mpg', 'curb-weight', 'horsepower', 'width' and 'city-mpg' (based on Table 1).

```
k_values = [1, 3, 5, 7, 9]
rmse_uni = {}
current_rmse = []
target_column = 'price'

for feature in continuous_numeric[0:-1]:
    for k in k_values:
        current_rmse.append(knn_train_test_uni(feature, target_column, normalized_cars, k))

    rmse_uni[feature] = current_rmse
    current_rmse = []
```

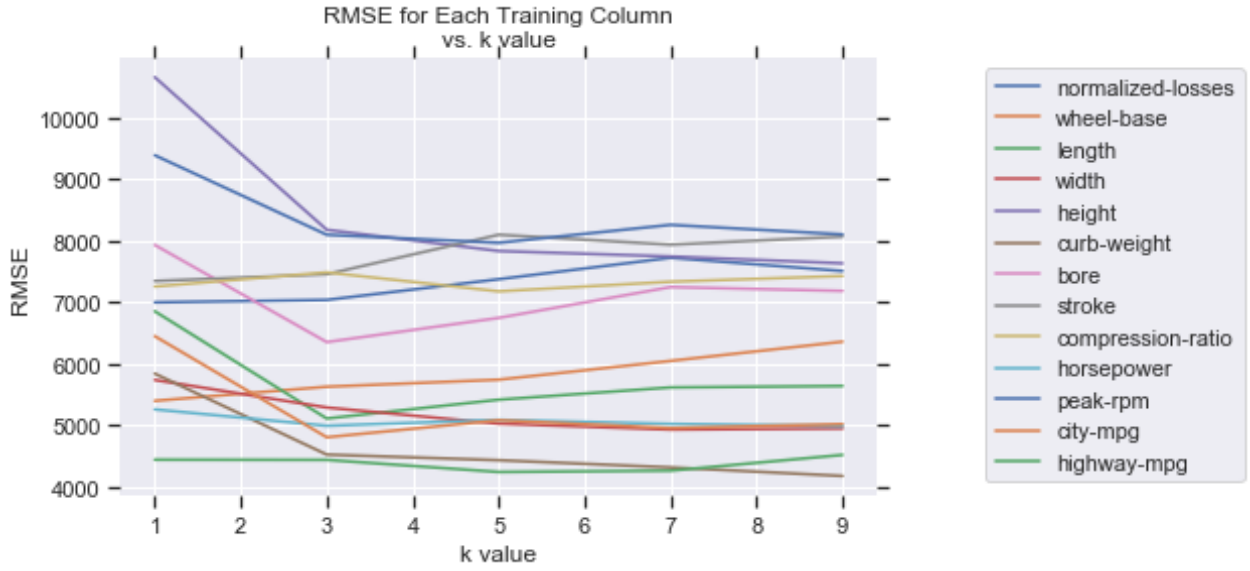

Fig. 2 RMSE Plotting for different k values

Table 1: List of RMSE values

Features	RMSE value
highway-mpg	4384.38
curb-weight	4661.53
Horsepower	5072.51
Width	5189.36
city-mpg	5264.74
Length	5729.05
wheel-base	5836.64

Bore	7092.32
normalized-losses	7328.33
compression-ratio	7335.78
Stroke	7779.82
peak-rpm	8361.30
Height	8407.70

KNN is a non-parametric, lazy learning algorithm. Non-parametric means no assumption is made for underlying data distribution. In KNN, K is the number of nearest neighbours. The number of neighbours is the core deciding factor. K is generally an odd number if the number of classes is 2. When K=1, then the algorithm is known as the nearest neighbour algorithm. In this case, we are classifying the car as '0' which denotes medium range cars and '1' which denotes higher range cars.

The confusion matrix was generated using the following code snippet and accuracy values are calculated. Table 2 enlists all the calculated values.

```
from sklearn.metrics import confusion_matrix
confusion_matrix(y_test, y_pred)
```

```
array([[34,  1],
       [ 0,  6]], dtype=int64)
```

Table 2: Accuracy Values for Different k Values

k Value	Accuracy (in %)
2	95.12
5	96.20
13	98.70
21	97.56

4. Conclusion

In this article, we have visualized the data using EDA (Exploratory Data Analysis), and predicted the price of car depending on various features of car like engine-size, horsepower etc. The benchmark dataset was chosen from UCI machine learning repository. Numeric values in the dataset was normalized. The price is predicted depending on other independent variables. RMSE values helped us to gauge which features are having great contribution in determining the price. The top five features appeared as 'highway-mpg', 'curb-weight',

'horsepower', 'width' and 'city-mpg'. The KNN model helped us to select the best features for price prediction. As the accuracy values were measured for different k values, it seemed that the performance is satisfactorily. At k = 13, the obtained accuracy was 98.7%.

As part of future scope, the dataset should be experimented with other classifiers to obtain better performance and better price prediction. Several other standard datasets from other repositories can also be explored. This kind of application of machine learning will enrich automobile or transportation domain.

References:

1. Lisa Ehrlinger, Elisa Rusz, Wolfram Wöß, "A Survey of Data Quality Measurement and Monitoring Tools", July 2019, arXiv:1907.08138.
2. Ralph Kimball, The Data Warehouse ETL Toolkit, Wiley, 2004. Florian Waas et al., On-Demand ELT Architecture for Right-Time BI: Extending the Vision, International Journal of Data Warehousing and Mining, April-June 2013, pp. 21-38.
3. Amar, L. A., Taha, A. A., & Mohamed, M. Y. (2020). Prediction of the final size for COVID-19 epidemic using machine learning: A case study of Egypt. Infectious Disease Modelling, 5, 622–634. https://doi.org/10.1016/j.idm.2020.08.008
4. Song, G., Rochas, J., El Beze, L. E., Huet, F., &Magoulès, F. (2016). K Nearest Neighbour Joins for Big Data on MapReduce: A Theoretical and Experimental Analysis. IEEE Transactions on Knowledge and Data Engineering, 28(9), pp. 2376–2392. https://doi.org/10.1109/TKDE.2016.2562627
5. Charles M. Judd, Gary H. McClelland, Carey S. Ryan, "Data Analysis: A Model Comparison Approach to Regression, ANOVA, and Beyond", 3rd edition, ISBN – 978-1-315-74413-1.
6. Elliott, V. (2018). Thinking about the coding process in qualitative data analysis. Qualitative Report, 23(11), 2850–2861.
7. John T. Behrens, "Principles and Procedures of Exploratory Data Analysis", Psychological Methods, 1997, Vol. 2, No. 2, pp.131-160.
8. Hadi, W., El-Khalili, N., AlNashashibi, M., Issa, G., &AlBanna, A. A. (2019). Application of data mining algorithms for improving stress prediction of automobile drivers: A case study in Jordan. Computers in Biology and Medicine, 114(September), 103474. https://doi.org/10.1016/j.compbiomed.2019.103474
9. Yaqi Li, Chun Yan, Wei Liu, Maozhen Li, "A Principle Component Analysis-based Random Forest with the Potential Nearest Neighbor Method for Automobile Insurance Fraud Identification", Applied Soft Computing, Volume 70, September 2018, Pages 1000-1009.