Advances and Applications in Mathematical Sciences
Volume 20, Issue 3, January 2021, Pages 367-375

COMPARATIVE ANALYSIS OF CAR SALES USING SUPERVISED ALGORITHMS

PRASHANT GUPTA, PRADUMN KUMAR, KUNDAN KUMAR and NIDHI SINGH

ABES Engineering College

Ghaziabad

Abstract

In this paper, we investigate the application of supervised machine learning techniques to predict the price of cars.

The predictions are based on historical data collected from Kaggle, an online verified repository. Different techniques like Linear regression, random forest, and decision trees have been used to make the predictions. We have then evaluated and compared the predictions so that we are able to find the algorithms and techniques which provide the best performances. All three methods provided comparable performance. In the future, we plan to use more techniques and methods.

1. Introduction

The prediction of car sales is an important and rewarding problem in current times. According to data obtained from the Government of India [1], the revenue generated from cars registered between 2009-2010 and 2015-16 has witnessed a spectacular increase of 34%. The number of cars in 2016 has reached 25,634,824. With a rise in new technologies and advancements, it is likely that sale of cars and the scope of this study will see a growth.

Kuiper [2] used a multivariate regression model to predict the price of 2005 General Motor (GM) cars. The results were satisfactory and can be repeated with accuracy. Support Vector Machines (SVM) were used by Listiani [3] to predict the sale of leased cars. It is concluded that SVM is

2010 Mathematics Subject Classification: 68W40, 91C20, 90C08.

Keywords: Car Sales Prediction, Data Analysis, Linear Regression, Decision Tree, Random Forest, Modeling.

Received May 20, 2020; Accepted July 31, 2020

more accurate in predicting prices as compared to the multiple linear regression when a very large data-set is under consideration. In medium or small-sized data-sets, Linear Regression suits well.

Noor and Sadaqat [4] also explored the idea of using Linear Regression to predict the sales of cars in place of SVM.

Accurately predicting the sale price of a car is a tedious yet rewarding action. Large numbers of features and records make the analysis very complex [5]. The said parameter is dependent on many factors that make up the characteristics list of the product. The most important ones are usually the mileage of the car, its make (and model), the origin of the car (the original country of the manufacturer), and its horsepower.

Unfortunately, if we take reality into consideration, a good majority of daily drivers are not aware of how efficient their vehicle is for every kilometer they drive. Other factors such as the type of fuel it uses, the interior style of the car, the acceleration, the capacity of its cylinders, its size, the weight of the car seem to be the most contributing factors. As we can see, the price depends on a large number of factors. In this work, we have considered only a medium-sized subset of the factors mentioned above.

2. Proposed Methodology

Data was collected from Kaggle [6]. We have tried to make sure that the data collected was as recent as possible. In most countries, seasonal patterns are somewhat of a minor problem only, as this does not affect the purchase or selling of cars.

Gonggi [7] proposed a new model based on artificial neural networks to forecast the residual value of private used cars. The main features used in this study were: mileage, manufacturer and estimated useful life. The model was satisfactorily accurate in nature.

In another car price prediction study [8], a neuro-fuzzy knowledge-based system was used. The following attributes were taken into consideration: brand, year of production and type of engine. The results were consistent with Linear Regression.

We have taken the following data about cars into consideration-model, horsepower, efficiency, year of manufacture, and price. Only cars which had their price listed were recorded. Because many of the columns were sparse they were removed. Thus, four Year Resale feature was removed. We made further modifications to our data-set by removing Model and Latest-Launch features as they had high cardinal values.

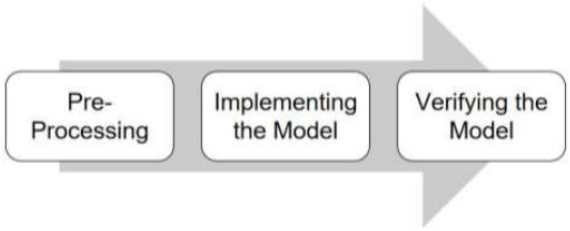

Figure 1. Flow of the Study.

2.1. Pre Modeling

The following methodology for pre-modeling has been found from Data Vedas [9]

i. Understand Data Types

ii. Find if data has Missing Values

iii. Find if data has outliers

iv. Understand the distribution of data

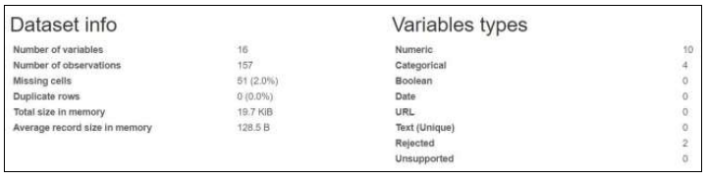

Figure 2. Types of variables.

Figure 3. Pre-processing required.

2.2. Modeling

- Y should follow a normal distribution

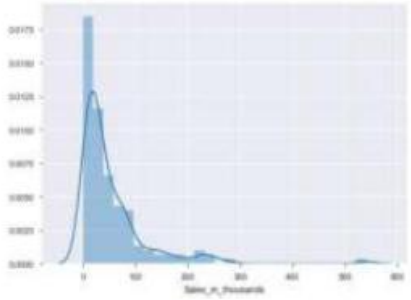

Figure 4. Before Normalization.

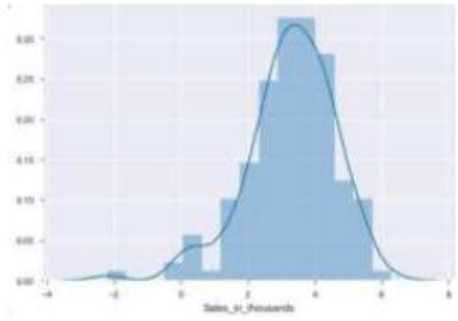

Figure 5. After Normalization.

- There should be correlation between X and Y

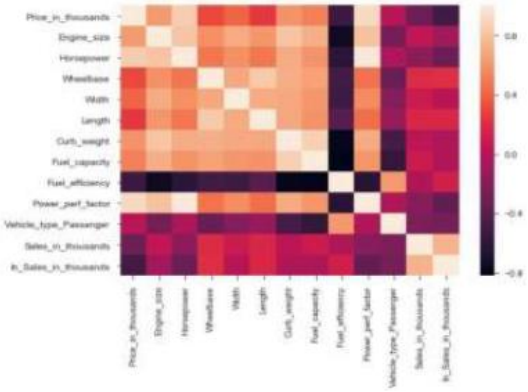

Figure 6. Correlation.

Feature Engineering can be of two types-

1. Feature Selection

2. Feature Reduction

Table 1. Tests for selecting Features.

F Regression (Univariate Regression)	F_values, p_values = f_regression(features, target)
Select K Best	KBest_Features = features.columns[selector.get_support()]
VIF (Variance Inflation Factor)	VIF = [variance_inflation_factor(features.values,i) for i range(features.shape[1])]

The data has to be split in train and test where on train dataset the model is built while on the test the stability of the model is tested. The data is split in 70-30 or 50-50 generally.

2.3. Model Implementation

2.3.1. Fitting the model on the test data

Table 2. Fitting various models.

Linear Regression	clf_linear.fit(train_X, train_y)
Random Forest	clf_rf.fit(train_X, train_y)
Decision Tree	clf_dt.fit(train_X, train_y)

2.3.2. Coming up with a mathematical equation that can be used for predictions.

2.4. Post Modeling

2.4.1. Model Validation

a. Apply the model on the test data-set

b. Get predictions

c. Comparing the model's performance on train v/s test.

2.4.2. Calculate Metrics (Model Evaluation)

We calculate the implemented model's performance using some metrics on the train as well as test data-set to check if the accuracy is high as well as similar for the data-set.

Table 3. Calculate performance of Model.

Mean Absolute Error	MAPE_train = np.mean(np.abs(train.Y - train.pred)/train.Y)
Root Mean Square Error	RMSE_train = metrics.mean_squared_error(train.Y , train.pred)

3. Implementation and Results

At first, data was made to go through pre-processing. All things that constitute towards making the data clean and trim for it to be able to pass through a model is pre-processing. Treatment for inaccurate data-types, handling missing values, and removing outliers was done in this stage. Missing values for numeric attributes were taken care of by filling them up with mean values of the data-set, while the categorical values were either dealt with removing the record or filling up with the mean of the data-set.

After confirming the data type accuracy, the data was ready to be checked for correlation. The data set correlation was handled by a heat map using python. The correlation gave us a good estimate of how many and how much if at all, any particular attribute affected the target variable. A brief check of other factors such as kurtosis and skewness were also done before implementing the modeling phase.

3.1. Linear Regression

The Linear Regression approach plots data on a graph and helps us to understand what is the best fit for the data. We can use that relationship and can predict further values based off of that relationship. In Univariate Linear Regression, variable (x) is input for the program and training is done on the basis of y. Data-set was modeled by keeping the price of the cars in the data-set as the target variable and other attributes as factors affecting it.

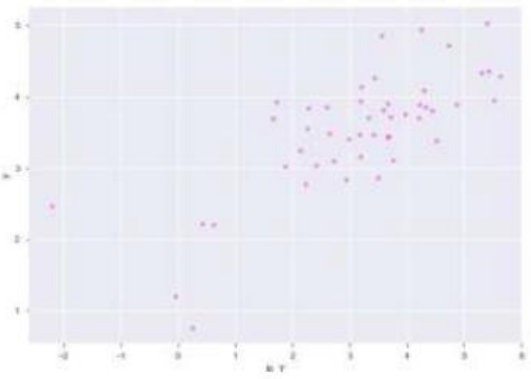

Figure 7. Result Comparison for Linear Regression.

3.2. Random Forest

Random forest is a type of classifier that incorporates different decision trees. Each decision tree provides a classification for input data and the random forest collects them and then chooses the best result as the optimal and final outcome. The random forest method uses multiple decision Trees. By taking the average of these trees, random forests tend to improve prediction. DIfferent Hyper Parameters can be used to obtain the best result.

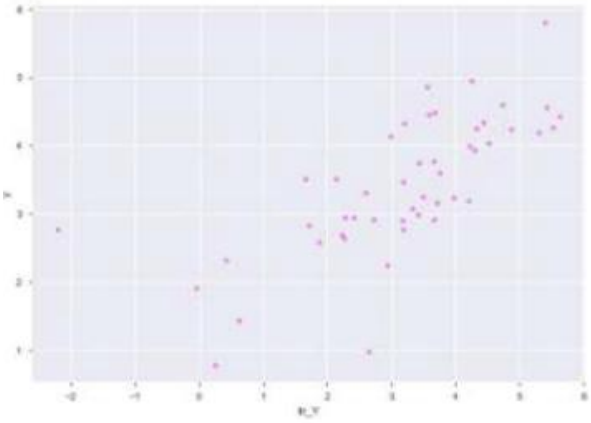

Figure 8. Result Comparison for Random Forest.

3.3. Decision Tree

A decision tree is a type of supervised learning algorithm. It works for both categorical and continuous input and output variables. In this

technique, we split the population or sample into two or more homogeneous sets (or subpopulations) based on the most significant splitter/differentiator in input variables. A decision tree or a classification tree is a tree in which each internal (nonleaf) node is labeled with an input feature. The arcs coming from a node labeled with a feature are labeled with each of the possible values of the feature.

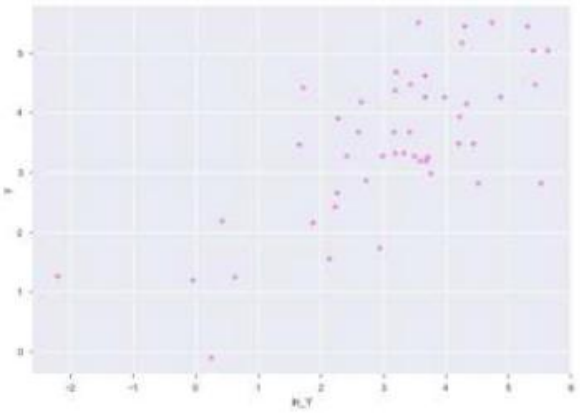

Figure 9. Result Comparison for Decision Tree.

After implementing and comparing the accuracy of three methods, we come to the following comparison-

Table 4. Accuracy of implemented models.

Model	Accuracy
Linear Regression	48.03
Random Forest	45.34
Decision Tree	44.67

Linear Regression provides the best accuracy at this point in time in the study. Random Forest and Decision Tree also contribute to the ongoing study path.

4. Conclusion and Future Scope

In our study, we studied and implemented three supervised machine learning algorithms to predict and estimate prices of cars. Data was split

into test and train and after necessary implementation steps, we found out that Linear Regression is able to most accurately predict the sales for this particular type of data-set. Any such data set that consists of industry-based most impactful features and a general mix of manufacturers will be accepted by the algorithm and a similar prediction can be replicated.

Furthermore, the algorithms can be combined and implemented to improve the accuracy of prediction as well. By giving more time and resources to this study, an approximate measurement can be made to list down top key parameters that affect the sales of cars for this particular type of data-set. Such a study flow will prove beneficial to car manufacturers and salesmen.

References

[1] State/ UT-wise Comparison of the fee received in National Permit Account from 2009-10 to 2015-16, retrieved from: https://data.gov.in/sector/transport

[2] S. Kuiper, Introduction to Multiple Regression: How Much Is Your Car Worth?, Journal of Statistics Education, 16(3) (2008).

[3] M. Listiani, Support Vector Regression Analysis for Price Prediction in a Car Leasing Application, Master Thesis, Hamburg University of Technology (2009).

[4] Kanwal Noor and Sadaqat Jan, Vehicle Price Prediction System using Machine Learning Techniques, (2017).

[5] Y. Singh, P. K. Bhatia and O. Sangwan, A Review of Studies on Machine Learning Techniques, International Journal of Computer Science and Security 1 (2007), 70-84.

[6] Car Sales Data-set, retrieved from: https://www.kaggle.com/gagandeep16/car-sales

[7] Gongqi, Shen et al, New Model for Residual Value Prediction of the Used Car Based on BP Neural Network and Nonlinear Curve Fit, 2011 Third International Conference on Measuring Technology and Mechatronics Automation 2 (2011), 682-685.

[8] J. D. Wu, C. C. Hsu and H. C. Chen, An expert system of price forecasting for used cars using adaptive neuro-fuzzy inference. Expert Systems with Applications 36(4) (2009), 7809-7817.

[9] Archish Kapil, Data Vedas. Archish Rai Kapil, 2018