4/25/2024

Norbert Shema

Course Project

IT 220 – Introduction to Data Analytics

# Project Description:

Through this project, we will use data extracted from Kaggle. The data in this file cardata.csv from company X contains different information on car brands. I split the data into two different datasets to be able to run different models. I.e. This project will be using the three excel files.

* Cardata
* cardata\_test
* cardata\_Train

The nature of the project is to predict the selling price of a car given different attributes. I will use a multiple linear regression model to predict the selling price. I will also use the KNN prediction model and weigh the best-fit model with fewer errors. I chose this project because I am recently looking to buy a car so knowing how different attributes included in this prediction would affect the price of an automobile. Also from a business standpoint, this project will help me understand the concepts of the two models I will be using, and I can apply them to any business idea I might want to predict in the future.

# Data Source and Description:

The data being used in this project was extracted from Kaggle.com, posted by (MALEKIAN, 2023). On this dataset, we have different attributes stated below:

**Car Name**

**Year**

**Owner**

**Present Price**

**Kms Driven**

**Fuel type.**

**Seller Type**

**Transmission**

**Selling Price**

The selling price and present price in this dataset represent values in thousands of dollars. In this dataset, we have **301** data values both numerical and categorical values such as the fuel type (Petrol, Diesel, or CNG) and the transmission (manual or automatic). On the file cardata.csv, on the second tab, we can see the results of the normalized values used in the prediction from WEKA.

A screenshot of a computer

Description automatically generated

Figure 1

The above figure 1 shows a sample of 29 top values of the dataset. We can also see this dataset's attributes from the first row.

A table of numbers and symbols

Description automatically generated

Figure 2

From Figure 2, We can see the normalized data from WEKA. We can see how Weka worked to change the attributes such as year. So that they can be easily read uniformly and worked with according to our data sets.

# Analysis of the Data:

A screenshot of a computer

Description automatically generated

Figure 3

Figure 3 shows the Weka Windows Explorer after we have imported the data set in Weka. We can see all the attributes displayed. I removed the Car name and the owner attributes as they won't help much to predict the selling price of a car.

**Linear regression model**

A screenshot of a computer

Description automatically generated

Figure 4

Figure 4 shows the output window after running the linear regression model on the dataset using Cross-validation of 10-folds.

We can see all the coefficients for the used attributes in figure 5.

A screenshot of a computer

Description automatically generated

Figure 5

In Figure 5, we can also see the categorical data like fuel type, seller type, and transmission, only use the coefficients as shown in Figure 5 otherwise the coefficient is zero. The model also predicts that if a fuel type is not diesel, then the coefficient would be zero. The same thing for seller type and transmission as all categorical data is true.

From Figure 4, we can see we have a great correlation coefficient of .9202 which shows a strong positive correlation between the variables. We can also see that the model prediction has an **RMSE of** **2.0142**.

Given the value, I would argue that the linear regression model does a great job predicting the selling price of the dataset given the other attributes.

The linear regression model equation would then be.

***Y=6.0475x + 40.2159x1 + (-3.5097) x2 + 1.8827x3 + 1.1635x4 +1.4344x5 + (-3.8721)***

**Y is the selling price.**

**X: year**

**X1: present price**

**X2: Kms driven.**

**X3: fuel type**

**X4: seller type**

**X5: Transmission**

A screenshot of a computer

Description automatically generated

Figure 6

In Figure 6 above we can see in the file Manual\_cardata how our model works towards predicting the selling price in Column I vs the actual value in Column H. In this step, I used the normalized data extracted from Weka and plugged in the linear regression model created. We can also see a new column for errors made for each predicted value.

**KNN prediction model**

I will use the K Nearest Neighbor (KNN) in this prediction model. To predict the selling price of a car giving all the attributes of it by comparing it to its KNN. The training set used in this model is the first 251 and the last 50 values have been set apart at the test values.

Experiment with different K values.

|  |  |
| --- | --- |
| K value | RMSE |
| 3 | 2.41 |
| 5 | 2.4827 |
| 7 | 2.6395 |
| 10 | 2.7291 |
| 11 | 2.7308 |
| 12 | 2.7127 |
| 15 | 2.7741 |
| 17 | 2.8829 |
| 20 | 3.0209 |

K=3 value yields the lowest RMSE. Now set the K value to 3 and test the 50 values in Weka.

A screenshot of a computer

Description automatically generated

Figure 7

Figure 7 shows the output window in Weka after I have used the supplied test set. We can also see the predicted values.

A screenshot of a computer

Description automatically generated

Figure 8

From Figure 8, We can see the remaining predicted values, and the **RMSE is 0.6051.**

A white screen with black text

Description automatically generated

Figure 9

Figure 9 shows a model created by using the MultilayerPerceptron prediction model. We can also see that the **RMSE is 1.4805.**

# Discussion and Conclusion:

In conclusion after running different models, Linear regression model, KKN, and Multilayer Perceptron on the data. KNN proves to be a better model that can be used to predict the selling price of a car given the different attributes. This is simply because it has the lowest RMSE of 0.6051, compared to 2.0142 linear regression model RMSE and 1.4805 for Multilayer Perceptron.

**Recommendations.**

Although the linear regression model was not the best model for the prediction of the selling price, but some recommendations can be drawn form the LR model analysis. According to the kilometers driven (x2= 40.2159) coefficient, an automobile's expected selling price decreases with increasing driving distance. Thus, consider highly as the vehicle's mileage affects the selling price with a bigger coefficient.

According to the coefficient for the year variable (x), the selling prices of newer cars are typically higher. Thus, you might want to think about purchasing a newer model if you want a good resale value. The current price, the selling price tends to rise in tandem with an increase in the current price, according to the present price coefficient (x1). Nonetheless, take into account whether the current price fits inside your spending plan and whether it supports the anticipated selling price.

# References

MALEKIAN, P. (2023). *Multiple linear regression on cars data*. Retrieved from Kaggle: https://www.kaggle.com/code/parnianmalekian/multiple-linear-regression-on-cars-data/comments