October 30, 2013

INTRODUCTION TO APPLIED MACHINE LEARNING

s1115104

Coursework 2

1. a

For the purpose of this section, only two attributes are considered – *price* and *engine power*. Visual representation of their relationship is presented in *Figure 1.1*. There appears to be a mild positive correlation between *price* and *engine power*, however, amount of data available for the first half of the horizontal axis is considerably greater than for second half and we would expect this relationship to skew the model. In practice, it is reasonable to think that engine power on its own is not sufficient to accurately predict the price of a car.

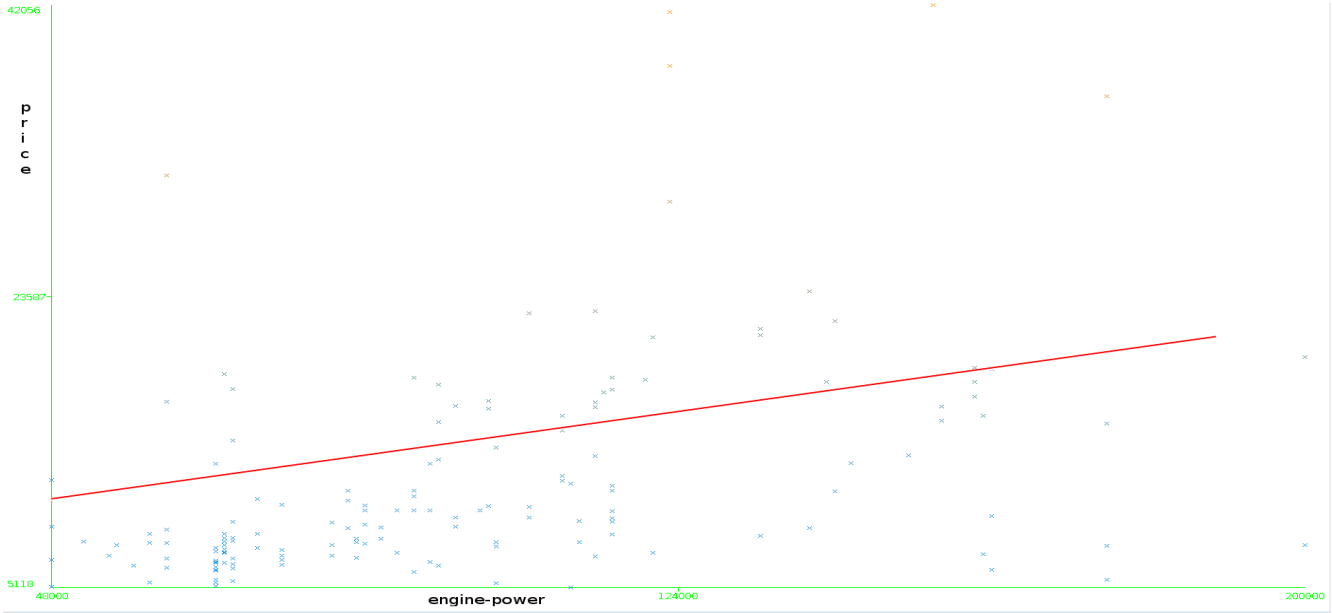

Figure 1.1: Engine power against price. Red line is the trend line (regression function)

1. b

Building a simple LR model in weka, we obtain a predictor function price in terms of engine power, $price(engine\_power) = 0.09 * engine-power + 3038.37$. Therefore, an increase in engine power by one unit will cause the price to increase by 3038.46 units.

1.c

Inspecting the model output, we obtain the following values:

Correlation Coefficient (CC)	Mean Absolute Error (MAE)	Root Mean Squared Error (RMSE)
0.4151	*3970.1023*	*6120.9373*

Table 1.1: Simple LR model

The coefficient of 0.09 suggests that price is not very strongly influenced by *engine power*. Indeed, we obtain *CC* of 0.4151, or around 41.51% of price increase can be explained by engine power, however, we cannot say anything about the rest. *MAE* measures how far off our predicted curve is from the actual points. *RMSE* is an alternative measure of how far off the prediction the points are.

1.d

Let us now compare our results with a simplest baseline model for regression. Building a simple LR model on *train_auto_partA_base* dataset. This dataset has all *engine power* values equal to one. The model built performs very poorly. We obtain the following values:

Correlation Coefficient (CC)	Mean Absolute Error (MAE)	Root Mean Squared Error (RMSE)
-0.1438	*4934.2762*	*6762.8527*

Table 1.2: Simple LR with $engine\_power = 1$

All of these values indicate a worse performance than before. In effect, our prediction is a constant of 11684.72, the mean of all prices.

2.a

Let us consider the dataset *train_auto_partB_numeric.arff*. Some attributes in this dataset are more useful at predicting prices than others. For example, *wheel-base*, *length* and *height* attributes appear to be fairly well correlated to price. On the other hand, torque and compression ratio appear to have almost no correlation to price and could be potentially removed from the dataset.

2.b

Building a multivariate LR model with numeric attributes only, we obtain the following values:

Correlation Coefficient (CC)	Mean Absolute Error (MAE)	Root Mean Squared Error (RMSE)
0.7307	*3121.3782*	*4837.2388*

Table 2.1: Multivariate LR

These are better results than for simple LR Model in the previous section, both MAE and RMSE have decreased while CC has increased, giving us better correlation.

2.c

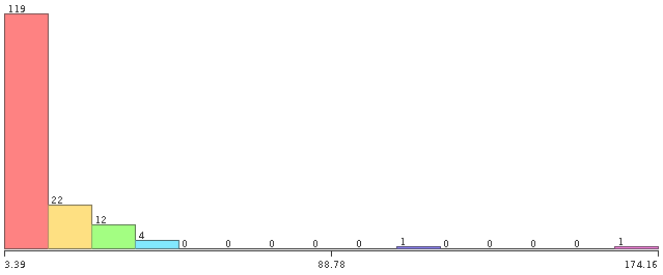

Figure 2.1(a): Engine-size default

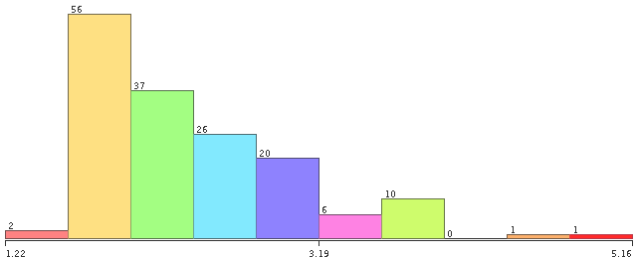

Figure 2.1(b): Engine size after logarithm function

Inspecting engine-size histogram in *Figure 3.1*, we would expect there to be issues with the model as majority of all values appears in the smallest quartile. We can use a logarithm function to transform the attribute and smoothen the values. Building the model again after engine size transformation, we obtain the following values:

Correlation Coefficient (CC)	Mean Absolute Error (MAE)	Root Mean Squared Error (RMSE)
0.8036	*2830.446*	*4029.8229*

Table 2.2: Multivariate LR with $log(engine\_size)$

This is, again, an improvement over the previous model. We have increased CC and lowered error levels further.

2.d

Interaction terms (IT) can be used to enhance the performance of a model. ITs attempt to capture real life properties of an object. For example, taking $size = width * height * length$ would attempt to model the size of a car. The following table summarizes results for various interaction terms involving engine_size.

	engine_power	peak_rpm	torque	mean-effective-pressure
CC	*0.7918*	*0.7237*	*0.5763*	*0.7339*
MAE	*2928.1583*	*3160.8833*	*4400.5068*	*3192.4095*
RMSE	*4246.8189*	*4951.6218*	*11354.0656*	*5229.9944*

*Table 2.3: Multivariate LR with interaction terms. Values are for engine_size * <column_name>*

Let us take *Table 2.1* as baseline for comparison, we can observe that the only case with better performance is of engine_power and engine_size. Mean-effective-pressure achieved higher CC than the baseline model, however, performed poorer in error measures.

2. e

Nominal attributes, contained in *train_auto_partB_full.arff* can be converted into binary with weka function *NominalToBinary*. We obtain a representation of values as binary attributes that can be directly used in multivariate LR model. The dataset cannot be used without the conversion as the LR model requires a binary relation with each attribute in order to use linear regression and return an answer in the form of a linear combination of a vector.

2. f

We can now classify the dataset. The *train_auto_partB_full.arff* dataset performs better than any of the other ones. We obtain the following results:

Correlation Coefficient (CC)	Mean Absolute Error (MAE)	Root Mean Squared Error (RMSE)
0.9295	*1741.2147*	*2542.4426*

Table 2.4: Multivariate LR with binarized attributes

CC has improved and now we have approximately 95% of the data correlated to the attributes of the model. If we consider results from Table 3.5 against all other results in this report, we can conclude that this model performs better than any other model examined. This is due to higher modularity of the attributes which provides flexibility for the LR model. On the other hand, by binarizing the dataset we obtain higher number of attributes and multivariate LR takes longer to compute.