MATRICULATION NUMBER – s1368177

Introductory Applied Machine Learning - Assignment 1.

PART 1

1.1.a)

Except for a few outliers, most of the points are scattered extremely close to each other near the origin of the plot i.e. (1,1). Distinct clusters of classes are not apparent. Hence, any discriminative approach such as KNN would be inefficient and constructing a decision boundary would be quite difficult.

1.1.b)

J48 Decision Tree Classifier had and accuracy of **~77.93**. Not only does this classifier deal with outliers very well, but also is capable of handling noisy data. For each discrete attribute, one test with outcomes as many as the number of distinct values of the attribute is considered. The performance percentage is high but not very high because trees divide the data into squares (only parallel to the axes) so building clusters around things means it has to split a lot to encompass clusters of data. Doing a lot of splits increases the chances of overfitting. Tall trees, like in this case, get pruned back so while one can build a cluster around some feature in the data, it might not survive the pruning process.

Naive Bayes Classifier had an accuracy of **~20.15%.** As mentioned in part a) about the clusters, the 2D Gaussians / marginal distribution for the classes would overlap which means that Naive Bays won't be able to distinguish between them effectively. Some of the newsgroups are very closely related. And,since the main difference between the classes is the amount of correlation, Naive Bayes independence assumption renders this classifier ineffective. Also, Naive Bayes is not equipped to deal with outliers, which skews or misleads the training process of the classifier algorithm.

As a baseline, we can consider a rudimentary classifier which classifies data points in a completely random manner. It should give an accuracy of almost 20% correct (rough approximation).

1.1.c)

Cleaning up the dataset with the following filter configuration:
weka.filters.unsupervised.instance.RemoveWithValues -S 100.0 -C first -L first-last -V

This removes the outliers whose frequency is greater than 100 and boosts the correctness of the Naive Bayes classifier to **~73.37%** from **~20.15%** since the presence of outliers was one of the reasons for Naive Bayes poor performance.

1.2.a)

InfoGainAttributeEval with the following method:

weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1

This ranks attributes in the increasing order of entropy. So the one with least entropy would be the most useful one. This is clearly evident in the results as team, hockey, rank, mac, and god are the five most useful attributes in the order and cancelled, bold, surface, warm, and sitting are the least useful ones. By visualising these against classes, one can notice that the most useful attributes are tightly and neatly clustered along the class axis whereas in the least useful ones they are spread vertically in the shape of a long bar throughout the length of the frequency axes.

1.2.b) After removing the twenty least useful attributes the performance of Naive Bayes Classification is increased to **~71.98%** from the **~20.15%** originally on the full dataset. This means that those twenty attributes were irrelevant, had high entropy and were only hindering the performance.

PART 2

2.1.a) Engine-power alone isn’t sufficient for predicting the price because there is no direct linear, inverse or exponential relationship visible just by looking at the plot. Furthermore, for most of engine-power points there is a cluster of vertically scattered price points, which means that for every major engine-power segment there are multiple prices which do not follow a pattern. To improve the performance of regression modelling, I would transform the data attempting to make coefficients more comparable and standardize based on the scale or range of the data so that coefficients are more directly interpreted and scaled, and remove outliers since outliers skew and mislead the training of the classifier.

2.1.b) The result from the regression model is as follows:

price = 0.0878 * engine-power + 3038.3671, When one more unit of engine power is added, the price goes up by the gradient of the regression model, **0.0878**, which in this case is the coefficient to the engine-power variable.

Engine-power is an important influential variable on price because if we plug in the mean value i.e. **98528.302** and multiply it by **0.0878** we get **8650.7849** which constitutes about **~74%** of the mean price. Therefore, it is an influential variable.

2.1.c)

Correlation coefficient - 0.405

Correlation Coefficient basically quantifies how well pairs of engine-power and price within their own distributions relate to each other which is useful in knowing how helpful is one variable in predicting the other.

Mean absolute error - 3999.335

Mean absolute error refers to the arithmetic mean of all the absolute errors without taking into account the direction of the error.

Root mean squared error – 6155.6971

Root mean squared error refers to the amount by which the values predicted by the classifier differ from the quantities being estimated i.e. variance of residuals.

A **correlation coefficient of 0.405 denotes a weak positive uphill relationship** between price and engine-power. That is, knowing engine-power alone does not give you enough information to predict price precisely. RMSE and MAE have a useful property of being in the same units as the response variable. Lower values of RMSE and MAE indicate a better fit, which is not the case here. Also the large difference (2156.3621) in both of these values denotes that there is a large variance in individual errors.

2.2.a) By looking at the plots, it is safe to assume that **torque** and **engine-size** have a higher correlation with price as compared to all the other attributes since they seem to follow a linear pattern with price. Also, **stroke, peak-rpm** and **mean-effective-pressure** have no correlation with price as the points are scattered everywhere in an arbitrarily. It is safe to assume that removal of these attributes can be done as they are not relevant for our model.

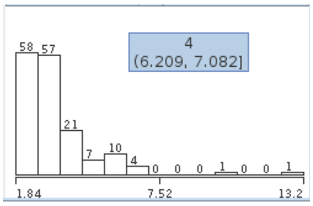
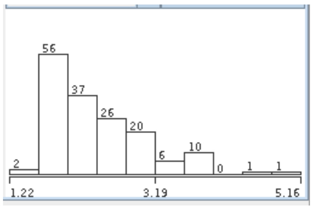
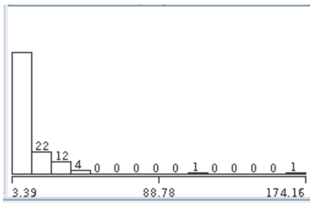
2.2.b)

Summary	Current	Previous
Correlation Coefficient	0.7307	0.405
Mean Absolute Error	3121.3782	3999.335
Root mean squared error	4837.2388	6155.6971

Comparison is shown in the above table where Current column is the one with all attributes from the new dataset and Previous, the one with only engine-power. **Correlation coefficient** has increased from **0.405 to 0.7307** which means that there is a strong positive uphill linear relationship between price and the data of all other attributes integrated. This is an optimistic sign as this model now is relatively more precise, if not perfect, at predicting the price. As for the MAE and RMSE, they both have also gone down significantly which means that the regression model is even better than before. The large difference between MAE and RMSE that was mentioned earlier has also gone down **to 1715.8606 from 2156.3621** which means that the large variance in individual errors has also reduced significantly.

2.2.c)

The histogram for engine-size is a bit extreme, in the sense that there is a large difference between the tallest and the smallest bars. This could skew the training for regression modelling and to prevent this from happening, we apply a transformation such as log or square root to normalize.

	Square-root transformation	Log transformation	NO transformation
Engine-size Histogram			
Correlation Coefficient	0.8062	0.821	0.7307
Mean Absolute error	2760.3064	2721.0033	3121.3782
Root mean squared error	4016.0449	3861.6063	4837.2388

From the above table, we can conclude that by transforming/normalising the attributes we can achieve a higher correlation, smaller individual errors and a smaller variance between individual errors. This happens because we ‘straighten’ the relationship into a linear one and that most statistical models use mean and normalising makes sure that the said mean is more accurately representative of the data.

2.2.d)

Attribute(s) Added	Expression	Correlation Coeff.	MAE	RMSE
None	None	0.7307	3121.3782	4837.2388
Drivetrain-size	Engine-size * wheel-base	0.6082	3519.2058	8683.3491
Engine-car-size	Engine-size * length	0.6574	3314.562	6999.4479
Engine-size-compression	Engine-size * Compression-ratio	0.8229	2750.0205	3931.2963
Engine-size-torque	Engine-size * torque	0.5763	4400.5068	11354.0656
Engine-size-stroke	Engine-size * stroke	0.7344	3155.6561	5643.3212
Engine-size-power	Engine-size * engine-power	0.7918	2928.1583	4246.8189
Engine-size-pressure	Engine-size * mean-effective-pressure	0.7339	3192.4095	5229.9944
Engine-size-losses	Engine-size * normalised-losses	0.7496	3068.2897	4656.3917

From the above table, we found that few new experimental interaction terms such as:

Engine-size-compression (0.8229), Engine-size-power (0.7918), Engine-size-losses (0.7496) and Engine-size-pressure (0.7393) had a higher correlation coefficient, lower MAE and RMSE than the original dataset and is a sign of improvement of the regression model. One interesting conclusion was that in 2.2.a) by looking at the plot, mean-effective-pressure seemed irrelevant, but combined with engine-size it has a positive and a relevant effect on our regression model.

NB. THE EXPERIMENT THAT FOLLOWS THIS LINE IS NOT A PART OF THE REQUIRED TASK SPECIFICATION:

Further experimentation which involved removal and addition of two attributes is as depicted in table below:

Attribute(s) added	Expressions	Correlation Coeff.	MAE	RMSE
None	None	0.7307	3121.3782	4837.2388
Engine-size-compression, Engine-size-pressure, Engine-size-power, Engine-size-losses	See above table	0.8447	2717.5517	3709.3197
Remove peak-rpm, stroke, bore	Removal			

By doing this experiment, some interesting results are obtained as shown in the table. We have managed to achieve a significantly higher correlation coefficient which suggests a strong positive uphill linear correlation and considerably low MAE and RMSE which is a sign of a good regression model relative to previous models. The difference between MAE and RMSE that was mentioned as reduced to **1715.8606** by including all attributes has been further lowered to **991.768** which is an even lower variance in individual errors.