Assignment 1 Feedback

1.1: Students generally forgot to use .info()/.describe()/.head: basically, they did not appropriately visualise the data

1.2: Students generally did not comment on .info() results, and did not comment on data anomalies (weird behaviour due to outliers)

1.3: Some students did not mention that there were outliers or extreme values of certain data points.

1.4: Students generally did not comment on the effect of outliers on the plot, and did not mention valid outlier detection methods (eg. histogram box-plot)

1.5: Students generally presented methods relying on assumed normality and filtered outliers based on positive and negative tails. But the data is not normally distributed and does not take negative values, (remember these are word counts)hence any such approach is wrong. The best approaches included sum of words, average frequencies or max-word threshold. Also, some students failed to appropriately visualise their techniques like Document length or Max value of feature vs. document index. Students generally did not mention that outlier removal significantly improved statistics.

1.6: Students did not use proper visualising techniques or forgot to present a clean and complete dataset. Students did not mention that cleaned data still has some high counts, meaning that a simple threshold might have taken some of the good documents away, and did not mention that outlier detection should be done on test set.

2.1: Students did not include either .info()/.describe()/head(): basically, they did not appropriately visualise the data.

2.2: Students did not give an accurate description of the Naive Bayes assumption. The most common wrong answer was that Naive Bayes assumes independence; instead Naive Bayes assumes **conditional** independence given the label. This is a very important distinction!

2.3: Some Students did not apply correctly the Naive Bayes assumption. Even though tek and ico are correlated in the graph, it does not tell us anything about whether it violates the NB assumption, again because of the difference between conditional and full independence!

2.4: Students generally did not describe how the dummy classifier works: i.e. it is wrong to just use the dummy classifier from SKLearn without an explanation. Also, we can do better than just random guessing (and there was an explicit post on Piazza about this) The purpose of a baseline is to have a simple method in order to compare it with more complicated ones – complicating it at this stage is not good practice. Also students generally did not report the class with highest prior probability.

2.5: Students generally did not analyse the number of instances in each class, hence it was more difficult to identify later that classes are relatively balanced. Some students did not compare Naive Bayes with the baseline, or did not answer whether accuracy is a reasonable metric (which it is in this case, given that the classes appear balanced).

2.6: Students generally did not give a satisfactory description of the best and worst classes. Moreover many students mentioned that the classifier confused class 2 with class 3 without mentioning an underlying reason or to which classes these labels correspond (for example

comp.sys.ibm.pc.hardware and comp.sys.mac.hardware) . This is why it is good to label plots: as it could be seen for example that classes 2 and 3, which are very related, are easily confused.

2.7: Some students retrained a new (dummy) classifier, when the question explicitly stated to use the same classifier trained in Q2.5 (and hence, the dummy should be the one trained on the same dataset, otherwise it is not a fair comparison). Some also failed to compare Naive Bayes with the baseline, or did not mention that performance of NB on the test set is acceptable given assumptions, and did not mention that there is no significant overfitting/ training..

2.8: Students generally reported wrong accuracy due to using the modified dataset in the previous question.

2.9: Students generally did not mention that Gaussian NB is not an appropriate model in this case, as we have counts (discrete non-negative) rather than continuous data!

Assignment 2 Feedback

In general, some marks were lost because plots were not labelled correctly, or floating point values where not correctly formatted.

1.1: Most students used the correct code (info/head/describe). However, when describing the dataset, most students failed to comment even on the data-types, or discuss the distribution (ranges) which should also have shown that there are no evident outliers.

1.2: Most students correctly calculated the correlation coefficients (although some did not compute it for the correct data). Students were also generally aware that **absolute** correlation was an indicator of ability of the feature to predict the price. However, in the last part, most students did not realise that features can be 'removed' for one of two reasons (both of which you should have mentioned): i.e. that either it has low-correlation with the price, or there is another feature with which it is highly correlated and hence gives little extra information…

2.1: The plot actually shows a weak correlation if anything.

2.2: Students did not in general give satisfactory answers as to why it is not conclusive whether the price-variable may be easy to be modelled. Basically, here we were looking for a discussion (1 sentence) on the fact that the assumption of the Linear-Regression Model is that the **residuals** are normally distributed with constant variance (conditioned on the independent variable). Since this is not displayed here, we cannot say conclusively whether we need to carry out the transformation. Also, some students failed to give an appropriate transformation (log or sqrt). Finally, most students did correctly notice that while some of the extreme values may appear as outliers, they could very well be very expensive cars (think Ferrari/Lamborghini etc…)

2.3: Most students got this right

2.4: The ideal baseline here was to just compute the mean on the **training set**: computing mean on all the data is wrong… we accepted answers where a linear model computed using the mean of both y and X was given…

2.5: Some students did not give the complete weights: here, we required both the bias (y-intercept) and first-order coefficient. With regards to comments, some students did not realise that one cannot just interpret the importance of a feature from a single weight: since this depends on scale… what you can do, is compute **relative** importance between **various** features IFF they are on the **same scale**! Other comments on R2 or other metrics were not required and useless in answering this question. Also, some students trained on the **entire** data, which biased the results and led to further mark deductions further on.

2.6: Most students correctly plotted the required data, although maybe not all plots were well labelled and included all that was asked for: that said, some plots completely missed the mark of what was required. The conclusion should have been however that it is not very conclusive which one is better (which then motivates the next set of plots).

2.7: Many students failed to answer parts of this question. In particular, you had to both discuss what the measures used (including the histogram) would be generally used for (basically, reading these from text-books/online) and then apply that analysis to the observations/data you have. For example, students should have interpreted the reason behind the negative R2 for the baseline (due to it being trained by mean on training data only!) or the fact that the histogram shows that models most often overestimate the price.

2.8: The deficiency in using a hold-out test has nothing to do with the fact that we are not reporting the values on unseen data: indeed we are. The problem is the small size of the data, which can cause two problems: imbalance in certain probabilities/distributions (note class imbalance is wrong as this is not a classification problem), and the fact that the testing set is too small to yield accurate statistics. Most students did not mention both.

2.9: Most students realised that the reduction in performance is due to the fact that the input data (engine-power) was constant and hence forcing the model to just learn the mean (much like a baseline). However, some marks were lost due to inappropriate documentation/explanation of the analysis.

3.1: Most students answered this correctly: some marks would have been deducted for incorrect formatting of output.

3.2: Some students plotted the wrong variable (meaning the rest of their analysis was wrong). It sufficed to say that the engine-size attribute has a highly-skewed relation with the price, which will be hard to model with a linear relationship.

3.3: Any transformation based on outlier removal is wrong. The correct way is to use a log/sqrt. Also, the student should have noticed that the significant performance increase, explained by both the above transformation as well as the fact that engine-size is a very important attribute to begin with (from our correlation study).

3.4: The answer here was to discuss that the scale of the original features has a key impact on the weights: to be able to compare features, they must be all on the same scale: ideally, this would involve a min-max transformation, although a standardscaler (normalisation) is also correct. Also, when showing output, given that we asked for the best three features, it was implied that you should sort your output to provide good presentation (besides good formatting)

3.5: In the selection process, you should have answered both questions: a way of choosing candidate features for higher-orders is to look at the scatter plots and pick those which appear highly non-linear: with respect to identifying the ideal order, cross-validation (hyper-parameter optimisation is also a good way of saying it) is required. Note that when adding the second order basis, you should make sure to: (a) keep also the original features, and (b) add also the cross-term length * engine-power. However, you should have noticed that the final result actually overfit the data (because we got worse generalisation performance)