Exploration

Factors impacting the reliability of Train-Test Split probing

Marco A. Wedemeyer - August 2020

A common rule of thumb used for determining a good train-test split is 80-20. In the last piece I looked at what the consequences were of changing this split and what generalisations could be made. The results indicated that there is a tradeoff between two sources of variance of RMSE when differing the random states: 1) test sets that are too small are more likely to be unrepresentative of the dataset as a whole and 2) test sets that are too large result in too little training data for the model to capture the underlying relationship. In this piece I want to explore different factors that impact the tradeoff between minimising mean RMSE and minimising the variability in reported RMSE[1].

The figures below show how the mean and the standard error of the reported RMSE change over the different test size proportions. The axes of the figures are limited in order to show regions of interest. The general shape of the graph showing mean over test set proportion follows an **L** shape. At high values of test set proportions the reported error is high as the models' are unable to learn a generalizable representation of the data. As these erroneous models are based on few instances chosen at random, the ways in which they are wrong differ greatly. This induces a very high standard error in the right tail. As the test set proportion decreases the mean error decreases until the limits set by the irreducible error are met. Further decreasing the test set is thus not able to decrease the RMSE and the value stabilizes forming the L shape. On the other hand, the smaller the test set becomes the higher the probability that the test set is less and less representative of the data set as a whole. The result is an increase in the standard error at the left tail, leading to a **U** shape.

The tradeoff that needs to be made is to have sufficient training data to learn the underlying relationship, in other words, reaching the long stabilized section of the L shape, and minimizing the chance of receiving an outlier RMSE value due to an unrepresentative test set. As the U shape of the standard error captures both of these opposing influences it is useful to mark the lowest point in the U as an indication for a good train-test split. These are visualized as black points in both graphs for each figure below.

[1] *Methodology* - The python machine learning package sklearn offers a way to generate synthetic data sets using its `make_regression` or `make_classification` functions. These two functions have parameters to impact the noise of the features (`noise`), the number of features (`n_features`), how many of the n features are informative (`n_informative`), and the sample size of the data set (`n_samples`). These attributes of data sets impact the tradeoff between mean RMSE and the standard error of RMSE in their own unique way.

Default data set parameters were set to improve comparability between the different trials. `noise` is set to 5 standard deviations, `n_features` and `n_informative` are set to 10 each and `n_samples` is set to 500. 20 different datasets were generated using different random states and averaged for each variation to smoothen the results.

Figure 1 Noise

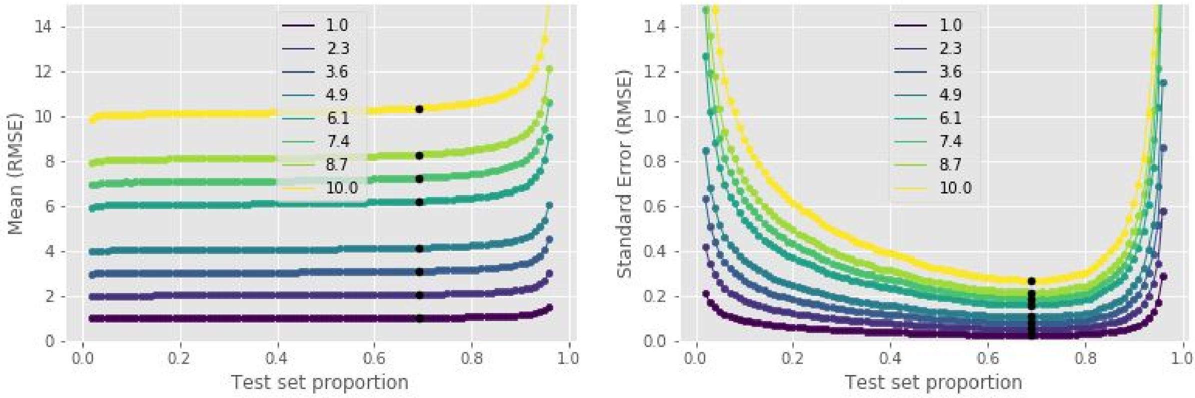


Figure 1 shows the impact of changing the `noise` parameter. As the irreducible error increases so does the mean RMSE. In general, the standard error also increases, however, more so on the left hand side of the U shape. With more noise in the data there is an increased chance of unrepresentative test sets emerging at larger test set sizes. Importantly, the optimal train-test split remains unchanged as the noise of the data varies.

Figure 2 Features

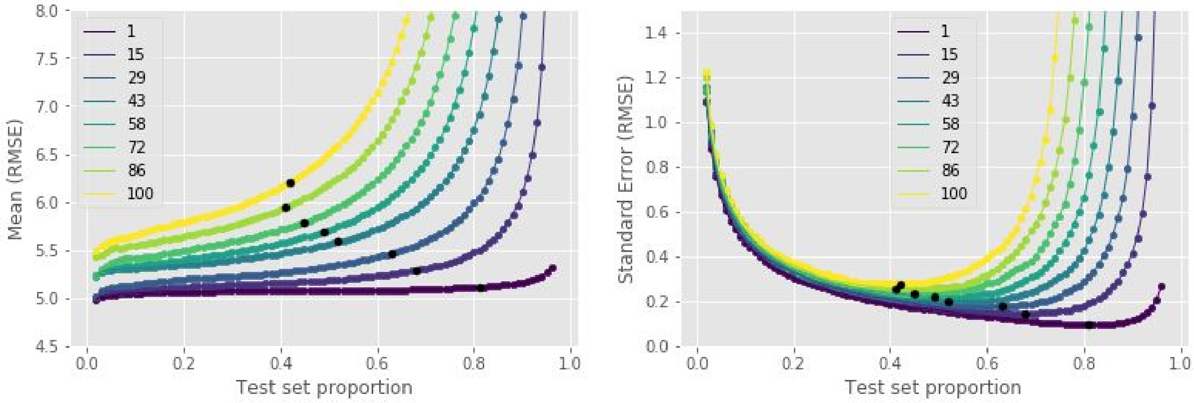


Figure 2.1 Modeling Mean RMSE mathematically

$\frac{-1}{ax+b}$

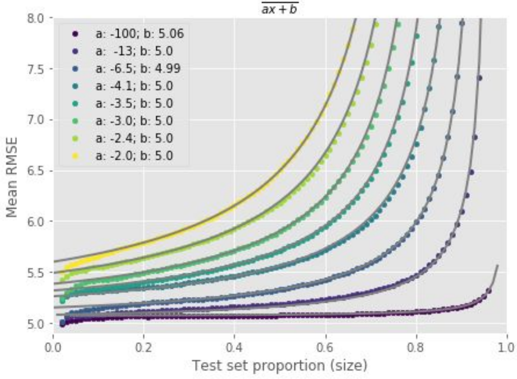


Figure 2 shows that as the number of features increases it becomes more difficult to model the underlying relationship. The mean RMSE increases drastically in the right tail as it requires larger and larger train set sizes to identify a generalizable representation of the data. The mean RMSE appears to follow a shifted -1/(ax+b) transformation of the test size. Figure 2.1 shows the data overlaid onto the mathematical estimates, which match very closely in the interval [5,8] of mean RMSE.

The values of the standard error in the right tail increase as the number of features increases, however, the left tail remains unchanged. The increased number of features changes the complexity of the problem and thus the amount of training data needed to model the relationship well, however, it does not affect the instability of a small test set.

Due to this one sided effect the optimal train-test split changes, as indicated by the movement of the black points.

Figure 3 Informative Features

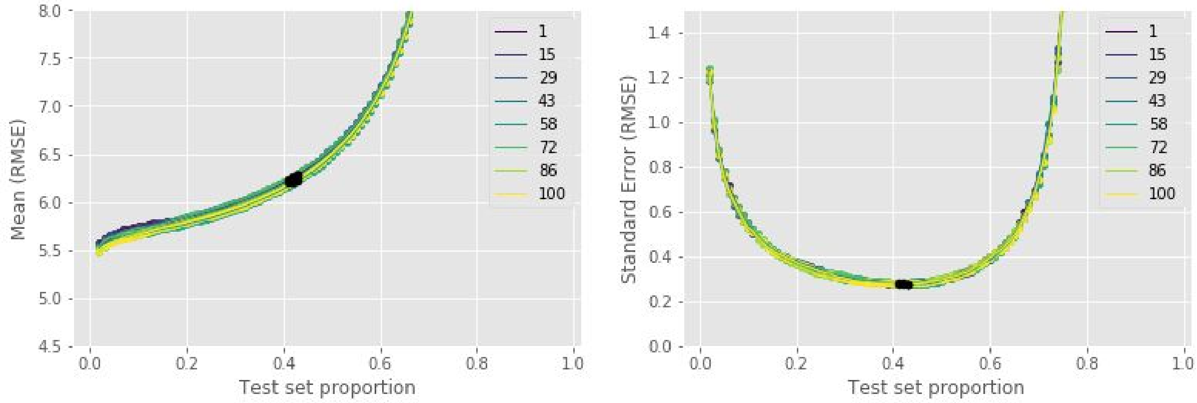


Figure 3 shows that changing the number of informative features, while keeping the number of total features the same results in no significant change in the mean RMSE or its standard error. The informativeness of a feature relates to its coefficient. Uninformative features have coefficients equal to zero, which the algorithm still needs to learn from the data. The complexity of the problem thus does not change and no effect is seen.

Figure 4 Sample Size

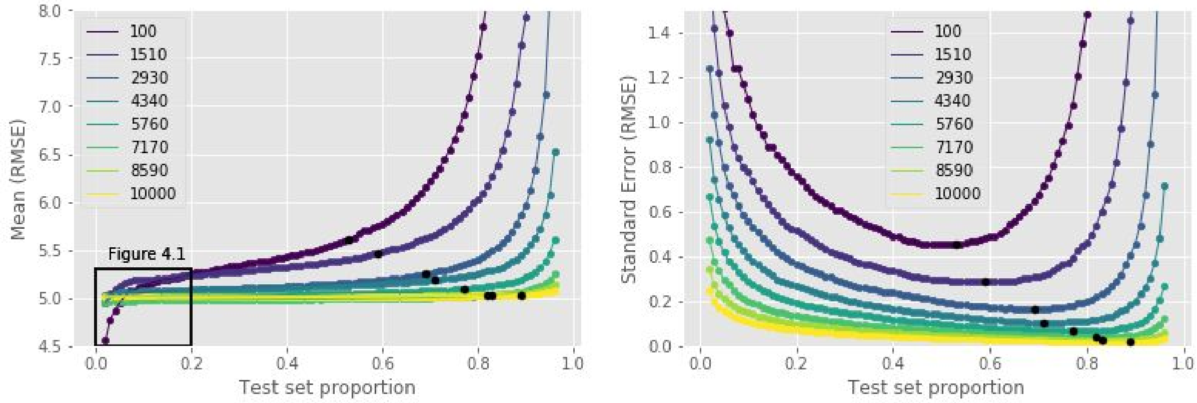


Figure 4 shows that by increasing the same size the mean RMSE and the standard error both decrease. As opposed to the other figures, the change in sample size is logarithmic. The way in which the mean RMSE decreases with changes in sample size follows a similar mathematical expression as for changes in the features but in reverse. Interestingly the mean RMSE dips below the noise parameter value of five for very small test sets at small sample sizes (e.g. 100), making the relationship appear closer to a third degree polynomial.

The mean RMSE should not dip below five, as this is the irreducible, unexplainable error. Yet, the smaller the sample size the larger the final dip in mean RMSE below 5. Figure 4.1 shows a close up view of the highlighted region in Figure 4. In Figure 4.2 the mean RMSE values were standardized to the value of each variation at a test size of 0.2. This graph shows clearly that the trend is not simply an outlier but a part of the relationship. I am yet unsure how the models are able to consistently outperform the added noise.

The standard error in Figure 4 declines as the sample size increases. With more data, the algorithm is better able to learn the underlying relationship. For instance, at a test set proportion of 0.97 and a sample size of 100 that leaves 3 instances for the training set, however, at a sample size of 10,000 the training set contains 300 instances. The same is true when applying this idea to the test set, which also reduces the standard error in the left tail. The reduction in standard error is however larger in the right tail, and thus the lowest standard error shifts towards the right hand side as sample sizes increase.

Figure 4.1

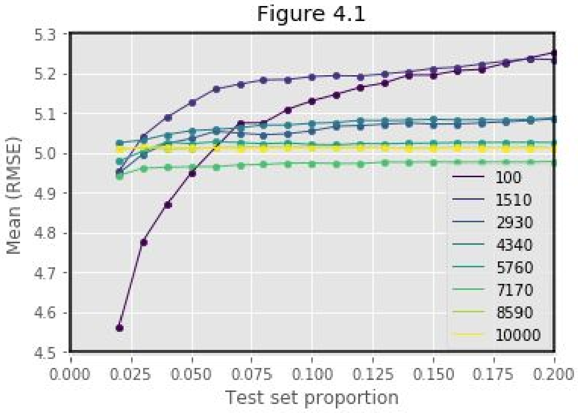
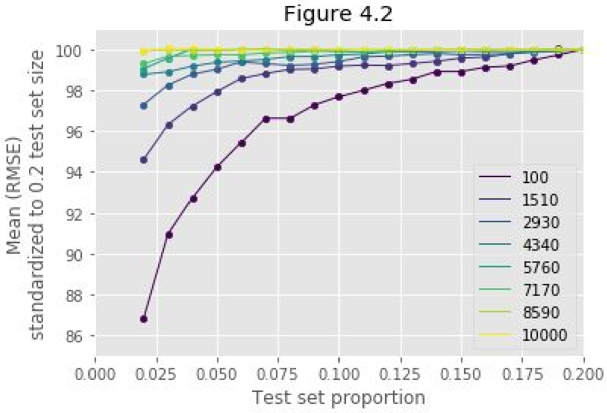


Figure 4.2



Conclusion - The takeaways from the presented results are that 1) although noise increases standard error it should not influence the choice in train-test split, 2) the number of total features influences the complexity of the problem, not the number of informative features, 3) the more features the data set has the larger the training set should be to reduce mean RMSE, and 4) the larger the sample size the larger the test set should be in order to reduce the standard error.

The numerical results shown in this exploration are to be taken with a grain of salt as they are derived for artificial data sets. The relationships between data set attributes should nonetheless hold and can meaningfully inform the choice of train-test splits. For further reading check out the article by Gianluca Malato on the same idea from a physics perspective