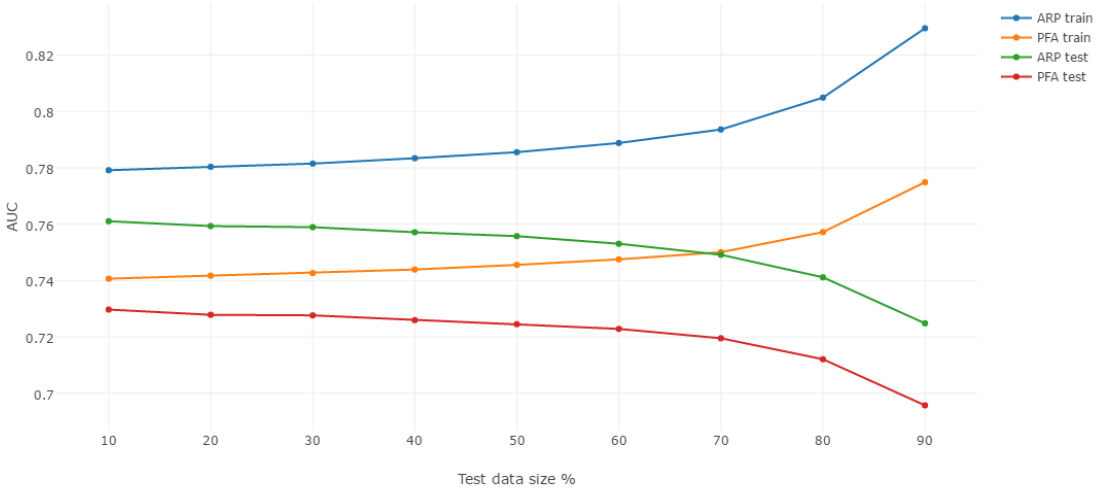

Figure 3.2: Over-fitting test of PFA and ARP models

fails. For example, as more polynomial terms are added to a linear regression, the greater the resulting model's complexity will be. In other words, bias has a negative first-order derivative in response to model complexity while variance has a positive slope.

Until this moment, it is unclear that how our new model works against over-fitting. Now considering over-fitting usually occurs when a model is excessively complex, such as having too many parameters relative to the number of observations, then it is possible for us to intentionally create scenarios that are to have prediction models over-fit in them. To be more specific, we are going to create a serious of training/testing splits on our data set, each with same amount of total data points but different testing data sizes range from 10% of all data to 90%, and run our models a numerous times with these data splits. We called this procedure the over-fitting test. By doing this test, we can observe when ARP and PFA start to over-fit and how large the errors are, as representations of the variance of our models.

Table 3.2 shows a plot of over-fitting tests of PFA and APR model. At each data split, we run both models 100 times with randomly selected testing and training data, and then measured the average AUC of these 100 runs. As we can see here, both ARP and PFA models share very similar "horn" sharp patterns that constructed by changes of testing and training performance using different data splits. ARP and PFA both show training performance better than testing performance across all data split settings. When testing data uses 10% of all data points, training models outperform testing models with small margins. As the size of testing data increases, training performance keeps increases while testing performance decreases. Take APR model for example, the AUC differences between training performance and testing performance start at $\delta_1 = 0.018$ then gradually increase to $\delta_2 = 0.100$. At that point, the training model has achieved an AUC of 0.829, but testing performance has decreased to 0.725 due to massive over-fitting. The absolute difference between two pairs of training and testing models at the beginning and the end of over-fitting test is $\epsilon = 0.082$. We believe that using δ_1 and ϵ is a reasonable measurement to quantify a model's degree of over-fitting, and can be used as a signal of variance. To our best knowledge, no prior work has formally utilized this information before, so we like to call this function as the *O-value* [1], and notated it as $O(\delta, \epsilon)$, so ARP model has a *O-value* at $O(0.018, 0.082)$. Respectively, PFA has a smaller *O-value*, which is $O(0.011, 0.064)$. Although ARP can be viewed better than PFA model when compared in the settings of 5-fold cross validation, however, when we conducting more closer investigation on model's variance, we see that two models perform neck and neck in general, and ARP model has ever slightly large variance in the measurement of *O-value*.

[1] The novelty and goodness of *O-value* are still being validated