Robust Models - 2

Spring 2020

Instructor: Ankit Shah, Ph.D.

Factors Affecting Model Performance

Model Performance

- Often companies get sub-optimal predictions due to reasons that are not related to the model
- One such aspect is the development of a model that answers a wrong question
- i.e., a question that the business really needs an answer for, is not what has been modeled

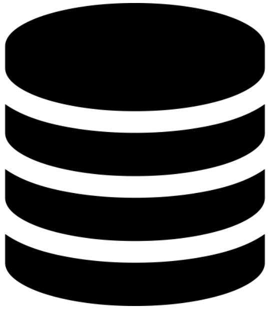
- Also, known as Type III errors

contacted via
email promotion

Contacted Customers

Customers that were not
reached (and new customers)

Stored Data



Information Collected
(Predictors)

- What time they were contacted?
- How many customers ordered?
- How much they ordered ($)?
- What they ordered?

Estimate (Response)

- Expected number of products that will sell

Decisions

- Which products to stock?
- How many quantities?

Noise: Measurement

- Data is collected through some process and there are some measurement errors that could occur. This is the noise/error associated with the measurement system.
  - $Y = f(X_1, X_2, \ldots, X_p) + \varepsilon$
  - The error term is independent of the predictor variables
    - e.g., measuring weight of an object on a weighing scale
    - If the scale has a systematic error – all the observations will be affected
  - Result: poor model performance
  - $\hat{Y} = \hat{f}(x_1, x_2, \ldots, x_p)$
  - $E(\{\hat{Y} - Y\}^2) = [\hat{f}(x_1, x_2, \ldots, x_p) - f(x_1, x_2, \ldots, x_p)]^2 + \mathrm{Var}(\varepsilon)$
  - Reducible error can be minimized by choosing a method that will generate better $\hat{f}$
  - Irreducible error is a constant and is not affected by the method used to produce $\hat{f}$

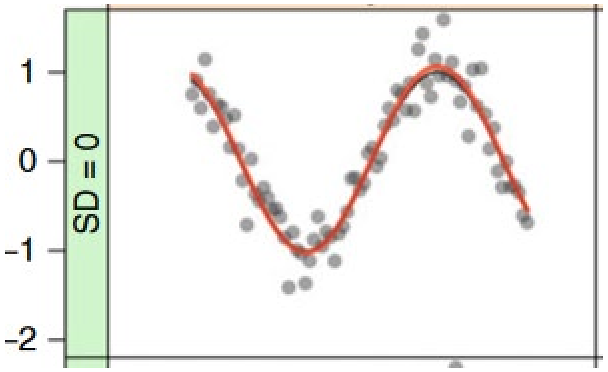
- The more the modeler understands the measurement system, the better is the understanding for the lower bound of the error

Given an extremely noisy system, will there be a significant difference in performance when one uses a highly flexible (complex) method vs a rigid (less complex) method?

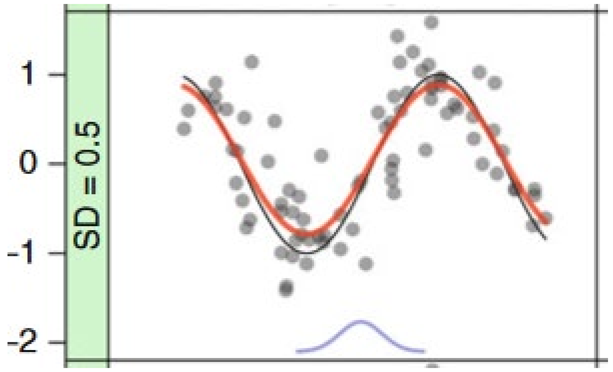
Noise: Measurement

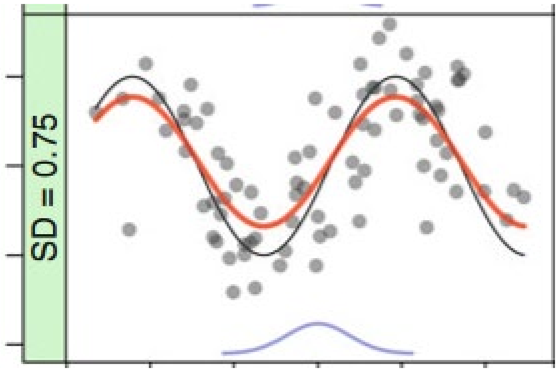
- Typically, it is assumed that there are no measurement errors with the predictors, but this is not always the case
  - An example is ratings conducting by humans
  - Adversarial techniques to manipulate predictor variables

- Data on the x-axis are evenly spaced values
- Response values are obtained by adding some normally distributed noise to the data



True fit (without noise)





As noise increases, the model performance decreases

Noise: Non-informative Predictors

- Another way noise can enter in your model is through attributes associated with the data points that have no relationship with the response variable
- There are certain methods that can eliminate or filter out such predictors, so they have no effect on the predictive performance of the model

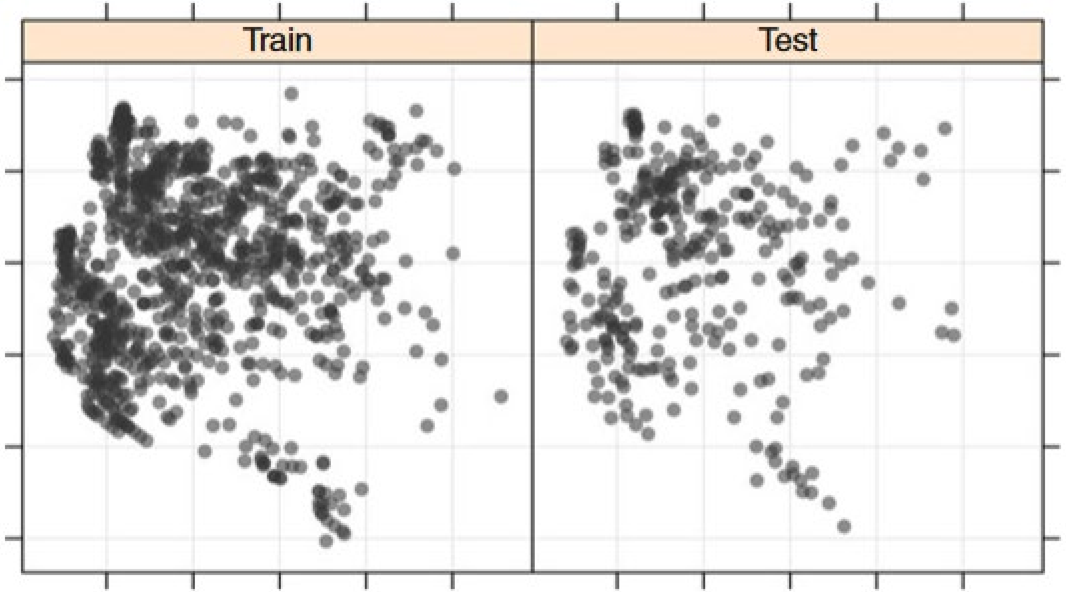
The Underlying Assumption

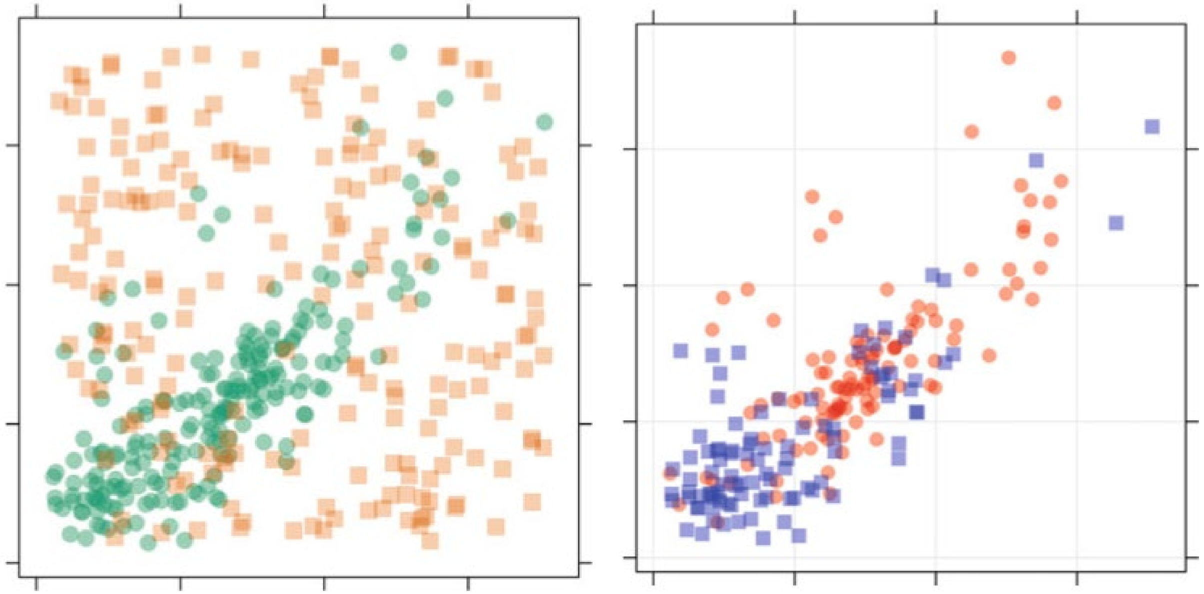
- An underlying assumption is that the mechanism that generated the training data set will continue to generate the new training samples
- Only in this event, we can be confident that the model we create will have good prediction accuracy for a new unseen data sample
- If this new sample is outside the range of the training data, what can be done?
  - Extrapolation
  - Such samples may not be trustworthy and will lead to poor predictions

- Is it possible to know if the underlying mechanism is same for both the test and train data sets?
- If there are a few number of predictors we can examine the scatter plots
- However, if the dimensionality increases, this will be inefficient

- The applicability domain of the model is the region of the predictor space where the model is expected to make accurate predictions

The Underlying Assumption





- If the training data and test data are generated from the same mechanism, then the projection of these data will overlap in the scatter plot
- However, if training data and test data are found in different parts of the scatter plot, then they might be coming from different mechanisms

Number of Samples

- It is assumed that the size of the training data set or the number of samples is directly related to the model's performance
  - If the data set is noisy, it minimizes any advantage that could be gained by a large number of samples
- One of the disadvantages is to add computational burden to train the model
- Imagine a single tree that exhaustively searches through all the samples and considers every predictor at a split to obtain the optimal splits at each level of the tree
- Now imagine the use of an ensemble technique, where we have many such trees
- There is a huge tradeoff between the model's performance and the computational burden
- This effect is compounded when the samples are from the same population
  - i.e., there is no new signal to learn/train the model

- Large data set is beneficial:
  - Samples contain information through the predictor space
  - Noise is minimal – the predictors and the response values
  - Samples are not similar
  - Computational burden is affordable