**What steps do you take when building a machine learning model?**

1. Define your problem
   1. What is your goal? How do you measure success?
2. Collect your data and prepare the data
   1. Wrangle up your dataset
   2. Prepare it for training (what features do you want to use?)
   3. Do some EDA, clean the data set of NaN, remove duplicates, correct errors, convert some features, normalization if necessary, remove biases
   4. Randomly split the data into training and test sets (with wafer data, randomization is important because the date the data is collected, matters.
3. Choose a model
   1. You want to consider prediction speed (is it a real time usage case?) and training speed (do you have to train it over and over again?), performance, ability to give you feature importance, what kinds of features can it deal with, supervised learning or unsupervised learning, regression/classification/clustering?, how interpretable do you need it to be and how complex?
4. Train the model
5. Tune your parameters
6. Evaluate the model
7. Make predictions

THIS IS MODEL building and diagnosis problem, there are many way to do it, according your statement I think the following will be easy. Assume the dependent variable is continuous and normal distributed as you stated

1.       Check the type of co factors, if some have a lot of missing value or typo, correct them or let them out. Then for continuous factors, check the person correlation coefficients,  if Pearson correlation near 1 or -1 among any two, one should be gone in multiple regression. For category factors, check independency of any two. if most count appear in the diagonal of the contingence table, one of the two category variables should be gone.

2.       Draw the scatter plots between the continuous outcome and each independent quantitative factor to see the association: if linear trend is shown, the factor is in, if non-linear effect is shown, transformation is needed. If no trend, like random, the independent factor could be out.

3.       Suppose the dependent variable should be independent. If the dependent variable is related to time, check the auto correlation, if autocorrelation exists,  time series modeling say Autoreg might be used.

4.       Do PCA analysis to see if there are still multicolinearity among the independent factors, if some eigen-value is near zero, you may drop one of them, or define a new factor (transformation) .

5.       If the sample size is large enough, say 10 (at least) times higher of the unknown parameter number, you can  do multiple regressions, you may use auto select option, such as forward backward or best, which will select independent factors for you.

6.       Number of parameters: interception: count 1, continuous factor, each counts 1, categorical factor with level of k, count as k-1.

7.       Check outliers by leverage or CooksD or Residual, if exists you may delete them or do both model with and without the outliers.

8.       Check normality of residuals from the multi variable regression,    if violated, do transformation, variance homogeneity exist: transformation on some independent factor, variance homogeneity doesn’t exist: transformation on dependent factor. Those may improve your model fitting.

9.       You may use Akaike’s information criterion or Bayesian information criterion or Mallows’ CP to decide how many factors should be in. Using them is better than comparison of  R2.

10.   Check interaction among the independent factors, the interaction among two quantitative predictors, means there is joint effect; the effect of one factor varies across the level of another factor. If the interaction between an quantitation factor and a category factor, say gender, it means the effect of the quantitative, slope,  is different between males and females.

Hope it helps.

Yuanzhang li