Machine Learning

You've collected your data, cleaned it up, wrangled it into shape and explored it. Now it's time to perform some in-depth data analysis using machine learning. This step depends on you and your mentor, but here are some suggestions to get you going.

1. How do you frame your main question as a machine learning problem? Is it a supervised or unsupervised problem? If it is supervised, is it a regression or a classification?

Original Research question:

What factors effect if a video becomes a trending video?

Research Question with Machine Learning in Mind:

Can we predict the number of views needed for a video to trend based on likes, dislikes, and comments? This problem is a supervised problem because the data is neatly arranged. Based on the variable of prediction, we frame the issue as a regression.

2. What are the main features (also called independent variables or predictors) that you'll use?

My independent variables are number of likes, dislikes, and comment count, and the output variable is views.

3. Which machine learning technique will you use?

The best model to use was linear regression to make a prediction for the number of views a video needs to trend based on the independent variables.

4. How will you evaluate the success of your machine learning technique? What metric will you use?

For this project, a training dataset will be created with 80% of the data from the original dataset, and the testing data set includes 20% of the data.

- The p-value, which determines whether the independent variables are significant. This will be used to determine a relationship between variables. The lower the p-value (ranging from 0 to 1), the more likely the model is more accurate at a certain level of confidence. The level of confidence is usually chosen at 95%.

- The R- squared value, which determines how “close” the points are to the regression line, is another metric that can be used to determine the validity of the model. The values range from 0 to 1, the closer the value is to 1, the better the prediction the model will give. If it is closer to 0, then the independent variable may need to be changed as it does not provide enough context or influence on the dependent variable and is therefore not useful for the model.

- Root Mean Square Error, the Root mean squared error is an average of the variance. This number is based on the units of your independent variable when deciding its fate. It can range from 0 to infinity. The smaller the number the better the model.