Data Analysis

Now, let's dive into the details of the process. Starting with the dataset itself. For the sake of our discussion, we limit the dataset into two kinds only: supervised and unsupervised with numerical and categorical data. There are many other forms of data, including text, image, etc. But we limit our discussion to these two only. For the supervised data, we will have classification and regression, while for unsupervised we will have clustering analysis. Again, our choices do not cover everything. We limit our choices for the most common and simpler ones.

The first step in our analysis is take a look at the dataset and determine what are we going to do with it. With supervised dataset, we are going to do classification algorithm if the dependent variable is categorical and regression otherwise. With unsupervised data, we are going to do clustering analysis. While classification and regression are basically finding a model that connect the independent and dependent variables. the clustering analysis in general is a method to find connection between data.

Now, once we have decided that we are going to do: classification, regression, or clustering, we then start to take a closer look at the dataset. We need to do some data cleaning. Handling missing data, if any, fixing format problems, etc.

Next, we turn to Exploratory Data Analysis (EDA). Note, when we made a decision whether we are going to do classification, regression, or clustering, we already started the EDA. Now we are going into more details, that is examining all variables and data more closely. What are we looking for? First, we are looking at individual variables. Are they really good behavior variables? Do they have outliers? How are their statistical behavior? Then, we examine the connection between variables without applying models, that is we are examining pair connections. We examine pair connection between independent variables and dependent variables. We also examine pair connection among independent variables. This is preliminary actions that are needed to obtain maximal model in the end of the process. Additionally, after we are finished with the final model we might want to go back to the EDA to have more explanation for the results or to get more ideas about how to improve the model.

Basically, EDA serves as (1) preparation for the machine learning part, (2) explanation, (3) inputs to improve the machine learning modeling.

Now we enter the application of machine learning. In case of supervised data, we want to extract a model that describe the connection between independent variables and dependent variables. Here, we should clarify the assumption that models from supervised machine learning served only as predictive tools. In our data analysis interpretation (i.e. explanation) is also important. So, we want to get interpretation and prediction. In case of unsupervised data, we want to do do clustering of data. For simplicity, we call the results of clustering as "model" because they are not necessarily exactly the same as original clustering existed in the original data.

Now we can discuss the product of our data analysis.

The products of data analysis are
(1) hidden pattern in the data, explanation about the underlying mechanism that govern the data, correctly construct predictive analysis, make predictions, and do prescription analysis,
(2) Find insights in every step of the way,
(3) Based on all the insights, write a story.

We can say that those three products are three different step of a process. The final product is the story. Now, what kind of story do we want? It's a story that integrates all the insights that we found into a meaningful whole. The purpose of the story is a guide for actions. By reading the story, the reader immediately understand the hidden pattern and the underlying mechanism that govern the data, the prediction, the needed prescription, and the necessary actions. The story put all insights into a single coherence presentation which help the reader understand everything and see how the recommended actions are indeed necessary.

One important aspect of the story is that it is meaningful because it is connected to the context and real problems in the situation where the data is coming from. While the statistics in the EDA, the algorithm, and the machine learning being used only care about the data itself (hence "data analysis"), the users have to implement the story to the real situation which is more than just "pure data".

The other aspect of the story is that while it has to be scientifically correct and rigorous, the readers have to understand it. It is conceivable that the readers are not data scientists, they are decision maker and people that has to do the real work.