TIG
Technology Integration Group®

TIG's Exploratory Data Analysis

Exploratory data analysis (EDA) is an approach we use to initially analyze our clients **data** sets to summarize the main characteristics, often with visual tools. This technique in Big Data and Advanced Analytics allows our Data Scientist and/or Analyst to build a statistical model primarily for seeing what the **data** can tell us beyond the formal modeling or hypothesis testing the task of predictive analysis - which would be the next step in our analytic process—post EDA.

Approach

TIG's Exploratory Data Analysis (EDA) is a method for data analysis that employs a variety of techniques to:

1. maximize insight into a data set
2. uncover underlying structure
3. extract important variables
4. detect outliers and anomalies
5. test underlying assumptions
6. develop parsimonious models
7. determine optimal factor settings

Focus

The EDA results can determine how a data analysis should be carried out. EDA is not identical to statistical graphics although the two terms are used almost interchangeably. Statistical graphics is a collection of techniques--all graphically based and all focusing on one data characterization aspect. EDA encompasses a larger setting in data analysis that postpones the usual assumptions about what kind of model the data follows. This gives a more direct view - allowing the data itself to reveal its underlying structure and model. EDA is not a mere collection of techniques. EDA is how we dissect a data set - what we look for; how we look; and how we interpret.

Techniques

Our EDA techniques are graphical in nature coupled with quantitative methods. The reason for the heavy reliance on graphical tools is that by its very nature the main role of EDA is to open-mindedly explore. Graphics gives the Data Scientist and Analysts' unparalleled power to do so, enticing the data to reveal its structural secrets, always ready to gain some new, often unsuspected, insight into the data. In combination with the natural pattern-recognition capabilities that we all possess, graphics provides unparalleled power to carry this out.

Primary and Secondary Goals

The primary goal of EDA is to maximize our Data Scientist and/or Analyst's insight into our client's data set as well as the underlying structure of a data set, while providing all of the specific items that an Analyst would want to extract such as:

1. a good-fitting, parsimonious model
2. a list of outliers
3. a sense of robustness of conclusions
4. estimates for parameters
5. uncertainties for those estimates
6. a ranked list of important factors
7. conclusions as to whether individual factors are statistically significant
8. optimal settings

Insight into the Data

Insight implies detecting and uncovering underlying structure in the data. Such underlying structure may not be encapsulated in the list of items above; such items serve as the specific targets of an analysis. The real insight and "feel" for a data set comes as the Data Scientist and/or Analyst judiciously probes and explores the various subtleties of the data with our clients and their SME's. The "feel" for the data comes almost exclusively from the application of various techniques, the collection of which serves as the window into the essence of the data. To get a "feel" for the data, it is not enough for the Data Scientist and/or Analyst to distinguish what is in the data; they must also recognize what is not in the data. The only way to successful accomplish this to draw on our own pattern-recognition and comparative abilities in the context of a series of judicious techniques applied to the data.