Independent EDA

- Start with summary statistics(min, max, range, std deviation, percentiles), missing values;
    - See if range, min, max values make sense. Could there be any data transformation problem.
    - What is the % of missing values? Could there be any systematic problem in getting data?
    - Watch out for default values; eg date 1-1-1970, many times people enter some random default values instead of missing values;
- Nominal variables;
    - See if factors have enough data points. If not, then consider combining factors;
- Next with histograms, frequency distributions--if skewed then perhaps transformations/binning would be a good idea; eg Salary is often positively/right skewed. Log transformation of salary typically fit better
- Box-plots again give sense of distribution, and outliers.

Relational EDA (This may also be used to explain your analysis to stakeholders)

- Correlations: See if some variables are highly correlated.
    - Think about it if you want to use them together. Perhaps factor reduction may be needed.
    - If using interactions eg, then variables are typically strongly correlated. Then demeaning helps to decrease correlations between interactions.
- Scatter plot matrix between response (continuous dependent) and predictor (continuous independent) variables
    - See if transformations may be needed;
    - Hint if some variables are key variables that need to be included
- Scatter plot (Plot by groups)—especially useful to check for moderation, ie say for lo-variable1, variable2 may have negative effect on dependent variable whereas for hi-variable1, variable2 may have positive effect on dependent variable.
- Line graphs: Sometimes data density may make it difficult to discern patterns in scatter plot, then line graphs can be good.
- Plot of means is good if response variable is continuous and predictors are categorical (if continuous then bin predictors).
- Boxplots (plot by groups) tells how response variable is distributed for various levels of categorical predictor.
- Cross-tabulation
    - works great to see relation between categorical variables. Continuous variables can be binned to conduct cross-tabulation.
    - Or numerical summary of continuous variable (group by categorical variable)
- Find good and bad examples by binning response variable into low, medium, high for example and then removing medium to have a stronger contrast.

Tree models:

- Response variable can be continuous or binary;
- Predictors can be nominal or continuous;
- Gives local patterns.
- The goal is to build a tree that uses the values of the input fields to create rules that result in leaves that are the purest in one of the target values.
- Good for exploration to identify important variables for other techniques;
- Sometimes it is easier to present results to management team as a decision-tree; Collection of rules make it easier to interpret for simpler trees;
- Can be used for classification, estimation.

- It handles null values better; Are not sensitive to outliers or skewed distributions.
- Choose the greatest complexity parameter whose estimated cross-validated error is still within a SE of the minimum possible cross-validated error. Hint: PlotCP; Try different models and validate using lift-charts;

Linear Regression:

- Response (dependent) variable is continuous;
- Predictors can be nominal or continuous.
- Assumption is that residuals (errors) are normally distributed
- Picks global level patterns; ie it cannot distinguish how one type of customers purchase behavior may be different. It assumes that the relationship between independent and dependent is same everywhere;
- Largely used for estimation eg customer life-time value, amount of purchase,
- OLS, closed form solution exists. It calculates an equation that minimizes the distance between the fitted line and all of the data points. OLS regression minimizes the sum of the squared residuals.
- R^2 is the percentage of the response variable variation that is explained by a linear model.
- Check residual plots to see if any systematic non-linearity has not been captured.

Logistic Regression:

- Response (dependent) variable is binary;
- Predictors can be nominal or continuous.
- If linear regression would be used
    - Predicted values would range beyond 0 and 1. Some value would be negative and some would be positive.
    - Errors will not be normally distributed;
- Odds ratio goes from 0 to infinity. Log of odds ratio goes from –infinity to +infinity; Hence errors are normally distributed.
- Picks global level patterns; ie it cannot distinguish how one type of customers purchase behavior may be different.
- Largely used for classification, something would happen or not, classifying good from bad, user would purchase or not.
- No closed form solution. MLE is numerical optimization.

Lift charts:

- Used to compare classification models—eg logistic regression, tree models;
- Can’t be used where response variable is continuous;