Steps of Data Exploration & Preparation

May 2018 Monday 21

Below are the steps involved to understand, clean and prepare your data for building model.

1. Variable Identification.
2. Univariate Analysis.
3. Bivariate Analysis.
4. Missing Value Treatment.
5. Outlier Treatment.
6. Variable Transformation.
7. Variable Creation.

We'll iterate over steps 4-7 multiple times before we comeup with our refined model.

May 2018 Tuesday 22

1. Variable Identification :

- Identify "predictor" (Inputs) and "Target" (output) variables
- Data types of the variables [char, Numeric]
- Category of the variables [category, Continuous]

2. Univariate Analysis :

- Univariate Analysis will depend on whether the variable is "categorical" or "continuous".

Sun	Mon	Tue	Wed	Thu	Fri	Sat
					1	2
10	11	12	13	14	15	16
24	25	26	27	28	29	30

Sun	Mon	Tue	Wed	Thu	Fri	Sat
3	4	5	6	7	8	9
17	18	19	20	21	22	23

JUNE 2018

23 May 2018 Wednesday

- Continuous Variable :

- We need to understand the "central tendency" or/and "spread" of the variable.

Central Tendency	Spread	Visualization
Mean	Range	Histogram
Median	Quartile	Boxplot
Mode	IQR	
Min	Variance	
Max	S.D.	
	Skewness	
	Kurtosis	

24 May 2018 Thursday Spread ≈ Measure of Dispersion

- Univariate Analysis is also used to highlight "missing and outlier values".

- Categorical Variable :

- We'll use "frequency table" to understand the distribution of each category.

- We can also read as percentage of values under each category.

- It can be measured using two metrics:

APRIL 2018

Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat
1	2	3	4	5	6	7	8	9	10	11	12	13	14
15	16	17	18	19	20	21	22	23	24	25	26	27	28
29	30												

"Count" and "Count %" against each category.

- "Bar Chart" can be used as visualization.

3. Bi-Variate Analysis :

1. finds out the relationship between two variable
2. We look for association and disassociation between variables at a pre-defined significance level.
3. We can perform bi-variate analysis for any combination of categorical and continuous variables.

4. The combination can be :
   - Categorical and Categorical
   - Categorical and Continuous
   - Continuous and Continuous

- Continuous and Continuous :

To find the relationship between two variables we used "Scatter Plot".

The pattern of scatter plot indicates the relationship between variables.

Sun	Mon	Tue	Wed	Thu	Fri	Sat
					1	2
10	11	12	13	14	15	16
24	25	26	27	28	29	30

Sun	Mon	Tue	Wed	Thu	Fri	Sat
3	4	5	6	7	8	9
17	18	19	20	21	22	23

JUNE 2018

27 May 2018 Sunday

The relationship can be linear and/or non-linear.

Scatter plot shows the relationship between variables, but does not indicate the strength of relationship amongst them.

Use "Correlation" to find the strength between variables.

It varies from +1 to -1. (-1 and +1)

-1 : perfect negative linear Correlation.
+1 : perfect positive linear Correlation.
0 : No Correlation

28 May 2018 Monday

- Categorical and Categorical

- Two-way table of Count and Count%
  - Rows represent Category of one Variable
  - Cols represent Category of other variable

- Stacked Column Chart : (more visual form of two-way table)

- Chi-Square Test

  - It is used to derive statistical significance between the variables.

APRIL 2018

Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat
1	2	3	4	5	6	7	8	9	10	11	12	13	14
15	16	17	18	19	20	21	22	23	24	25	26	27	28
29	30												

- Chi-Square is based on the - difference between the expected and observed frequencies in one or more category in two-way table.

- It returns the probability for the computed chi-square distribution with the degree of freedom.

  . Probability of 0 : both cate var are dependent

  . Probability of 1 : both cate var are Independent

  . Probability <0.05 : It indicates that the relation ship between the variables is

significant at 95 % confidence.

- Categorical and Continuous :

  - We can use box-plots for each level of categorical variables, to explore the relationship between categorical and continuous variables.

  - To look at the statistical significance, we can perform

    • Z-test, T-test or ANOVA .

Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat
					1	2	3	4	5	6	7	8	9
10	11	12	13	14	15	16	17	18	19	20	21	22	23
24	25	26	27	28	29	30							

- Z-test / T-test : Either test assess whether mean of two groups are statistically different from each other or not.

- If the probability of Z is small, then the difference of two average is more significant.

- The T-test is very similar to Z-test, but it is used when no. of obs. for both categories < 30.

- ANOVA : It assesses whether the average of more than two groups is statistically different.

4. MISSING VALUE TREATMENT :

1) Why Missing Values treatment is required ?

- Missing data in the training set can reduce the power / fit of a model.
- It can lead to biased model because we have not analyzed the behaviour and relationship with other variables correctly.
- It can lead to wrong prediction or classification.

2) Why my data has Missing Values ?

• Missing Values can occur at two steps :

(a) Data Extraction :

i) It is possible that there can be problems with extraction process.

ii) In such cases, we should double check the data with data guardians.

iii) Some hashing procedures can also be used.

(b) Data Collection :

- These error occur at time of data collection and are harder to correct.
- Categorized into four groups

• Missing completely at random :

- This is a case when the probability of missing value/variable is same for all observations.

Example : respondants of data collection process decide that they will declare their earnings after tossing a fair coin. If a head occurs, respondent declares his/her earnings and vice-versa. Here, each observation has equal chance of missing value.

• Missing at random :

- This is a case when variable is missing at random and missing ratio varies for different values/level of other input variables

Example : We are collecting data for age and female

MAY 2018

Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat
		1	2	3	4	5	6	7	8	9	10	11	12
13	14	15	16	17	18	19	20	21	22	23	24	25	26
27	28	29	30	31									

has higher missing value compare to male.

- Missing that depends on unobserved predictors:

  - This is a case when the missing ratio varies for different values / level of other input variables.

  - This is the case when the missing values are not normal; and are related to the unobserved input variable.

Example: In a medical study, if a particular diagnostic causes discomfort, then

there is higher chance of drop out of the study. This missing value is not at random unless we have included "discomfort" as an input variable for all patients.

- Missing that depends on the missing value itself:

  - This is the case when ø the probability of missing value is directly correlated with missing value itself.

Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat
1	2	3	4	5	6	7	8	9	10	11	12	13	14
15	16	17	18	19	20	21	22	23	24	25	26	27	28
29	30	31											

JULY 2018