Sachet Misra, Feixiang Xi, Yufei Wu – IIT (Data Science)

Data Science Practicum

Data Processing and Plans – Presentation I

Inferences – to be found

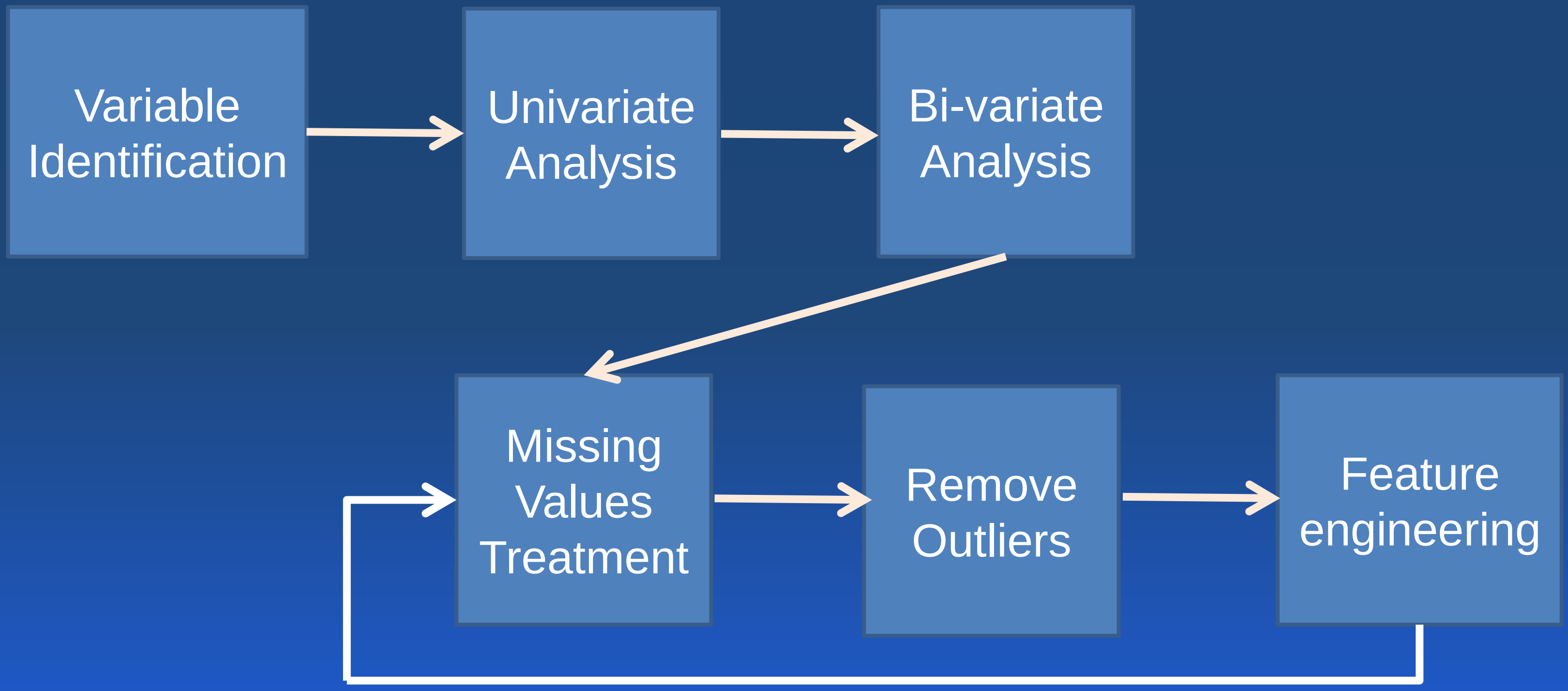
- which problem to solve
- an overlapping time period
- the building with the highest and most reliable amount of data
- the most efficient method of data storage

Define a Problem

- Gives us Direction
- Helps us use the above techniques efficiently
- Some techniques are iterative (model dependant)

The Approach

A Diagram



Variable Identification

Univariate Analysis

Bi-variate Analysis

Missing Values Treatment

Remove Outliers

Feature engineering

Check with model (multiple times)

Techniques

- Variable Identification
- Univariate analysis
  - Continuous
  - Categorical
- Bi-variate analysis
  - Categorical & Categorical variables
  - Categorical & Continuous
- Missing Values Treatment
  - why data could have missing values
  - methods of treating missing values
- Remove Outliers
- Feature creation / Feature engineering (Variable transformation / creation)

Variable Identification

- to help us find the data required and not required data
- also find the input variables (Predictor) and the output variables (Target)
- Target can be better identified when we define the problem

Univariate Analysis

- Continuous
- Categorical

Bi-Variate Analysis

Categorical & Categorical

- relationship between two categorical values
  - two-way table
  - stacked column chart
  - chi-square test
- use this method while solving a similar problem of a different building
- can also find if there are some categories of data we can create

Bi-Variate Analysis

Categorical & Continuous

- can draw box plots
- use tests like ANOVA, Z-test, T-test to find the significant variables
- significant variables reduces efficiency of the models but decreases the time required for computation required for the model
- the efficiency reduction might be negligible

Missing values treatment

(why data could have missing values)

- Data Extraction
- Data Collection
  - missing completely at random (probability of the missing value of a variable is same for all observations)
  - missing at random (variables missing at random - ratio varies for other input variables)
  - missing that depends on unobserved predictors (not random but depend on the unobserved input variable)
  - missing that depends on the missing value itself (there is a direct correlation between the missing value and the probability of the missing value)

Missing values treatment

(methods of treating missing values)

- Deletion
- Replacement using Statistical Methods (mean, median, mode imputation)
- KKN Imputation

Outliers
(Detection)

- Univariate & Multivariate Model
- Causes (artificial or natural)
  - data entry errors
  - measurement errors
  - experimental error
  - intentional outlier
  - data processing error
  - sampling error
  - natural outlier

Outliers
(Removal)

- Deleting Observations
- Transforming and Binning values
  - Natural log
  - Binning
  - Decision tree algorithm
- Imputing
  - statistical methods - mean, median, mode
- Treat Separately

Feature creation / Feature engineering

Will Talk about this while building models

Tools

- Microsoft's data profiling task editor
- Azure Data Factory Service
  - Azure SQL Database
  - Azure SQL Data Warehouse
  - SQL Server Database
  - Supports Hive, Pig, R Script and Hadoop
- OpenRefine – Google
- Open Source Data Quality and Profiling, etc.

Timeline

- 2 Weeks
  - profiling,
  - Selecting / Understanding the problem
  - Getting to speed with the technology to be used
- 3 Weeks
  - modelling and getting data ready
  - prediction
- week 10 finalising the model

thank you!