greatlearning

Guidelines for PGPDSE FT Capstone Project – Interim Report

Industry Review

- Industry Review – Current practices, Background Research
- Literature Survey - Publications, Application, past and undergoing research

Dataset and Domain

- Data Dictionary
- Variable categorization (count of numeric and categorical)
- Pre Processing Data Analysis (count of missing/ null values, redundant columns, etc.)
- Alternate sources of data that can supplement the core dataset (at least 2-3 columns)
- Project Justification - Project Statement, Complexity involved, Project Outcome – Commercial, Academic or Social value

Data Exploration (EDA)

- Relationship between variables
- Check for
    - multi-collinearity
    - distribution of variables
    - presence of outliers and its treatment
    - statistical significance of variables
    - class imbalance and its treatment

Feature Engineering

- Whether any transformations required
- Scaling the data
- Feature selection
- Dimensionality reduction

Assumptions

- Check for the assumptions to be satisfied for each of the models in

- Regression – SLR, Multiple Linear Regression, Logistic Regression
- Classification – Decision Tree, Random Forest, SVM, Bagged and boosted models
- Clustering – PCA (multi-collinearity), K-Means (presence of outliers, scaling, conversion to numerical etc.)

---------- **Interim Presentation Checkpoint**----------