**Guidelines for PGP-DS Capstone Project**

**Industry Review**

- Industry Review – Current practices, Background Research

- Literature Survey - Publications, Application, past and undergoing research

**Data set 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, and Project Outcome –Commercial, Academic or Social value

**Data Exploration (EDA)**

- Relationship between variables

- Check for

- Multi-co linearity

- 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-co linearity), K-Means (presence of outliers, scaling, conversion to numerical, etc.)

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

**Model building**

- Split the data to train and test.

- Start with a simple model which satisfies all the above assumptions based on your dataset.

- Check for bias and variance errors.

- To improve the performance, try cross-validation, ensemble models, hyper parameter tuning, grid search

**Evaluation of model**

- Regression – RMSE, R-Squared value,

- Classification – Classification report with precision, recall, F1-score, Support, AUC, etc.

- Clustering – Inertia value

- Comparison of different models built and discussion of the same

- Time taken for the inferences/ predictions

**Business Recommendations & Future enhancements**

- How to improve data collection, processing, and model accuracy?

- Commercial value/ Social value / Research value

- Recommendations based on insights

----------------------------- **Final Presentation Checkpoint**----------------------------------------------------------

**Dashboard**

- EDA – Correlation matrix, pair plots, box blots distribution plots

- Model

- Model Parameters

- Visualization of performance of the model with varying parameters

- Visualization of model Metrics

- Testing outcome

- Failure cases and explanation for the same

- Most successful and obvious cases

- Border cases

----------------------------- **Final Submission Checkpoint**----------------------------------------------------------