Data Science Life Cycle Project:

A general data science lifecycle process includes the use of machine learning algorithms and statistical practices that result in better prediction models. Some of the most common data science steps involved in the entire process are data extraction, preparation, cleansing, modelling, and evaluation etc.

1. Exploratory data analysis (EDA)
2. Feature Engineering
3. Feature Selection
4. Model Training + Hypo Parameter tuning
5. AO ops ==> CI / CD ==> Deployment
6. Model retraining approach

1. Exploratory data analysis (EDA):

Exploratory data analysis (EDA) is used by data scientists to analyse and investigate data sets and summarize their main characteristics, often employing data visualization methods.

1. Analysing the Data
2. Visualizing the data
3. Understanding the data
4.descriptive stats

2. Feature Engineering:

Feature engineering or feature extraction or feature discovery is the process of using domain knowledge to extract features (characteristics, properties, attributes) from raw data.[1] The motivation is to use these extra features to improve the quality of results from a machine learning process, compared with supplying only the raw data to the machine learning process.

1. Handling Missing values
2. Handling Categorical features
3. Handling Outlier
4. Handling Imbalanced data
5. Feature Transformation
6. Feature Extraction (PCA)
7. Creating derived feature

3.Feature Selection:

Feature selection is the process of reducing the number of input variables when developing a predictive model. It is desirable to reduce the number of input variables to both reduce the computational cost of modelling and, in some cases, to improve the performance of the model.

1. Correlation
2. VIF

3 Random Forest --> feature Important

4. Extra tree
5. Chi Squire
6. Annova test

4. Model training:

Model training is the phase in the data science development lifecycle where practitioners try to fit the best combination of weights and bias to a machine learning algorithm to minimize a loss function over the prediction range.

1. Train with every Model
2. Randomised search CV
3. Grid search CV
4. TPoT --> Generic Algorithm

5. AI ops

==>CI / CD pipeline ==> Deployment (AWS, GCP, Azure)

6. Model Retraining Approach.

Model retraining involves lifting and shifting the batch training code defined at development time into an automated workflow. You should abstract feature selection, model parameters, and other configurable pipeline parameters as input variables of the retraining pipeline.