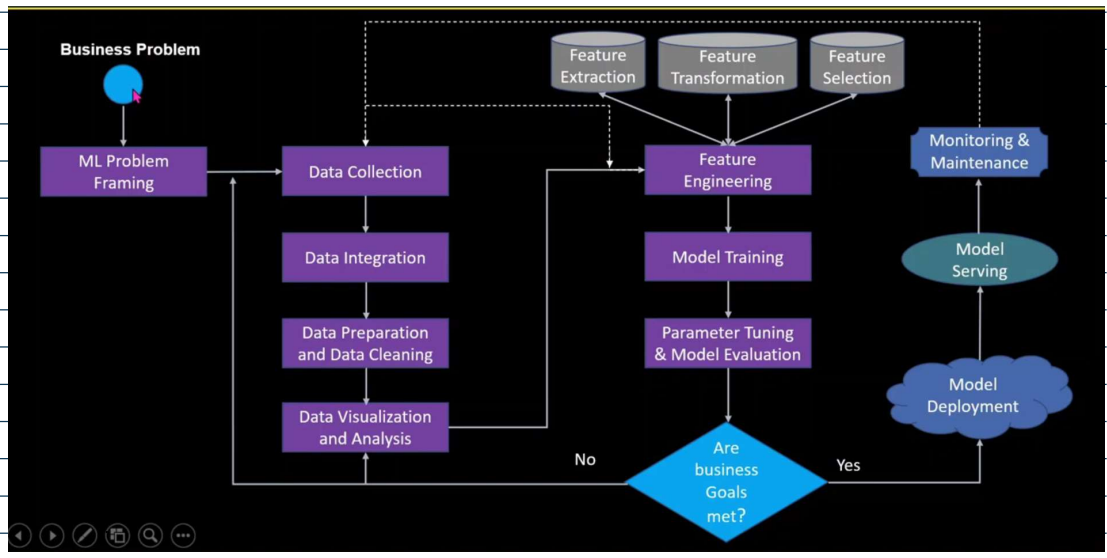
MLDLC

Monday, 4 March 2024 6:16 pm

1. Frame the problem
   - where will we get data from
   - Offline or Batch mode
   - what are we trying to solve
   - who are our target audience
   - cost
   - team size
   - How will end product look
   - Is the model supervised or unsupervised
   - What type of algorithms will help us

2. Gathering Data
   - CSV
   - API
   - web scraping
   - Database → Data warehouse [ETL]
   - Spark Clusters

3. Data preprocessing
   - Remove duplicates
   - Remove missing values
   - outliers
   - scaling



4. EDA - study input & output relation, experiment with data visulization, Univariate / Bivariate / Multivariate
   Imbalanced data.

5. Feature Engineering / Selection

6. Model Training, Evaluation & selection
   - Performance metrics
   - Hyperparameter tuning

7. Model Deployment

8. Testing (Beta testing) A/B testing

9. Optimize.