Overview

1. Clarify Requirements
2. How the ML system fits into the overal product backend
3. Data Related Activites
4. Model Related Activities
5. Scaling

Details

1. Clarify Requirements
   - What is the goal? Any secondary goal?
     - e.g. for CTR - maximizing the number of clicks is the primary goal. A secondary goa
   - Ask questions about the scale of the system - how many users, how much content?
2. How the ML system fits into the overall product backend
   - Think/draw a very simple diagram with input/output line between system backend and ML
3. Data Related Activites
   - Data Explore - whats the dataset looks like?
   - Understand different features and their relationship with the target
     - Is the data balanced? If not do you need oversampling/undersampling?
     - Is there a missing value (not an issue for tree-based models)
     - Is there an unexpected value for one/more data columns? How do you know if its a
   - Feature Importance - partial dependency plot, SHAP values, dataschool video (reference)
   - (ML Pipeline: Data Ingestion) Think of Data ingestion services/storage
   - (ML Pipeline: Data Preparation) Feature Engineering - encoding categorical features, emb
   - (ML Pipeline - Data Segregation) Data split - train set, validation set, test set
4. Model Related Activities
   - (ML Pipeline - Model Train and Evaluation) Build a simple model (XGBoost or NN)
     - How to select a model? Assuming its a Neural Network
       1. NLP/Sequence Model
          - start: LSTM with 2 hidden layers
          - see if 3 layers help,
          - improve: check if Attention based model can help
       2. Image Models - (Don't care right now)
       3. Other
          - start: Fully connected NN with 2 hidden layers
          - Improve: problem specific
   - (ML Pipeline - Model Train and Evaluation) What are the different hyperparameters (HPO)
   - (ML Pipeline - Model Train and Evaluation) Once the simple model is built, do a bias-variai overfitting vs underfitting and based on whether overfit or underfit, you need different appr
   - Draw the ML pipeline (reference #3)
   - Model Debug (reference #1)
   - Model Deployment (reference#3)
   - (ML Pipeline: Performance Monitoring) Metrics
     - AUC, F1, MSE, Accuracy, NDCG for ranking problems etc.
     - When to use which metrics?
5. Scaling