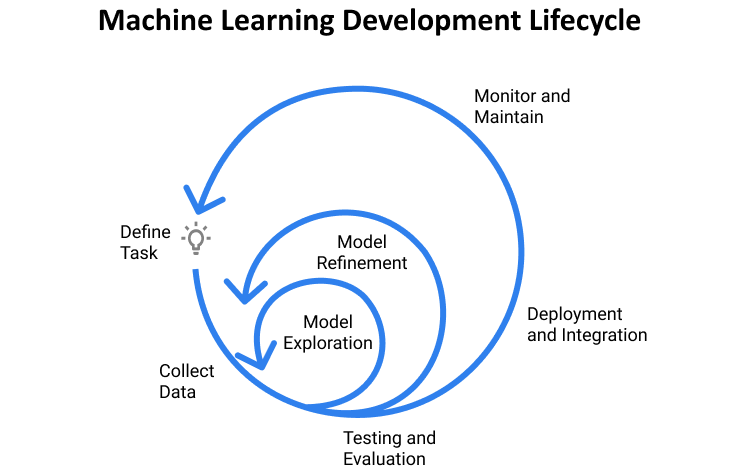
ML Sys backend question



Business Problem :

There is a huge problem and opportunity in automating the ERP process, one biggest pain points are extracting line items from invoices and converting them to existing DB entries in an ERP system. There are other business use apart from the automation wrt to invoice, understanding what people buy, spend analytic, mapping to GL code which department pays etc.

Line item extractions involve following major work.

* Table Detection
* Table Structure
* Functional analysis - understanding each column table headers
* Table classification - what table is what? Price table or

Data :

There are lot of online data,

* problem is they are not real world,
* they are also overly over segmented
* They

We labeled ourselves, then used self supervision and snorkel to bootstrap the training data to create more example

Quality check for them

Data Augmentation

MinMax, RandAugement

**Model - visual**

Baseline - FastRCNN -

Model 1 - Cascade rcnn → However, the commonly used threshold values �, typically � = 0.5, establish quite a loose requirement for positives. [Cascade RCNN | Hasty.ai](https://hasty.ai/docs/mp-wiki/model-architectures/cascade-rcnn)

Model 2 - multimodal

Model 3- DERT - was another option for us to use but atleast on our evaluation set it was performing worse, also could be number of examples that we had for bipartite loss function and transformer based model to scale and learn was less

Loss functions :

Text it was

Fasttext for learning embedding and for classification based classification

→ easily to train with n-gram modeling, loss function was negative sampling over hierarchical softmax as there were a vocabulary was large amount of infrequent word

→ classify as it as date, amount description, as table or not

→

Structure was graph cnn

→ Structure was two methods

→ table, table column, table row, table cell.

→ Graph CNN - Each token became, vertex, edge become interactions “column” is coonection, row is connection, converting it into vertex features, [hand crafted, is it bold, has line , spacing etc]

→

**Metrics**

Classification : precision and recall

Image processing - Average precision and average precision at 50

**Training and** Validation

1. Check for underfitting
2. Check for overfitting
3. Check for bias
4. Check for variance

**Production**

DataModel

Monitoring

—-------

Start with the problem statement and end with the business impact.

1. What is that you were trying to solve.
2. Cover the data collection and data cleaning aspects. Any specific data related issues that you faced like class imbalance or significant null values etc.
3. Explain your choice of ML model - why did you use RandomForest and not SVM.
4. Tell how did you improve the model.
5. What were the final results. Which metrics were you tracking.
6. How did you optimize the model.
7. How did you deploy the model in production.
8. What was the business impact.