- Accuracy: Questions related to training accuracy and/or convergence

- Evaluation
  - Evaluation method selection: Question related to the problems in the usage of APIs for doing validation.
  - Visualizating model learning: Questions about visualizing behavior of model to get a better understanding of the training process & the effects of evaluation on the change of loss function & accuracy

- Hyper-Parameter Tuning
  - Improving models performance
  - Tuning strategy selection: Question about choosing among APIs for different tuning methodologies

2022-05-11 Tuning parameter selection: Discussions related to the selection of parameters for tuning 2022-05-12

- Prediction
  - After model trained and evaluated, the model is used to predict new input data.
  - Prediction accuracy: Questions related to prediction accuracy
  - Model reuse: Developers might have difficulty in reusing existing models with their own datasets
  - Robustness: Question about stability of models with slight change, possibly noise, in the datasets

Manual Labelling

- Participant Training: - Participants were provided with the classification
  - Training session was conducted.

- Labelling Effort: - Eeah participant gave each question one of the labels from top-level categories namely: Non-ML, Data Preperation, Modelling, Trainäng, Evaluaton, Tuning, Prediction.
  - Then assigned a Sbcategory

Reconciling Results: - Moderator ~~creat~~ compared labels collected.
- 177 labelled question were disputed and discussed for resolution.
- Measured the inter-rater agreements using Cohen's kappa coefficiant (k) which measures the observed levels of agreement between raters of a particular set of nominal values and corrects for agreements that would appear by chance.
- Fleiss coefficient which is widely used for finding IRR between more than 2 raters

Threats to validity

→ mitigated bias by using more than one author

- Internal validity: manual labelling can be biased.
  - mitigate bias of missing posts by using particular library
  - classifying top level catgoires bias was mitigated with PHD students studying subset of posts + 3 ML experts
  - ML expertise of the raters can affect the manual labeling. To mitigate the threat theused raters with ~~pro~~ expertise in ML

Hilroy

External validity: • low quality posts and chronological order of posts. To eliminate quality threat we studied only the posts that have the tag of the relevant library. Only kept post with score ¿=5
• Chronological order of posts can introduce threat as some older posts
• expertise of programmer asking question.

~~RQ1: DIFF~~

RQ1: Difficult Stages

Most Difficult Stage
• Model creation is the most challenging (yet critical) in ML pipeline, especially for libraries supporting distributed ML on clusters like Mahout and MLlib

Data Preparation.
• Data preperation, especially data adaptation, is the second most difficult stage in ML pipeline

Type mismatch

RQ: Nature of Problems
• Type mismatches appear in most ML libraries

Shape mismatch.
• Shape mismatch problem appear frequently in deep learning libraries. Keras is an outlier in this subcategory with 5.5% of posts.

Data Cleaning
· Most libraries have problems in data cleaning

Model creation.
· Problems that are both inherent to ML, & specific to design choices in the library
· Model creation for deep neural networks is difficult as well
· problems in model creation due to the dependency of the model on multiple files
· Having several components complicats matters.

Error Exception.
· Questions on exceptions/errors are prevalent
·

Parameter Selection.
· Parameter selection can be difficult in all ML libraries

Loss function selection
· Choice of loss function is difficult in deep learning libraries.

Training accuracy
· Abstract ML libraries have higher percentage of question about training time accuracy and convergence.

Tuning parameter Selection
· Scikit-learn has more difficulty in hyperparameter tuning compared to other libraries

~~Correlation~~

Correlation between libraries

- Weka, H2O, Scikit-learn, MLlib from a Strong correlated group with correlation coefficient greater than 0.84 between the pairs indicating that these libraries have similar problem in all the ML stages
- Deep learning libraries, Torch, Keras, Theano, and Tensorflow form another group with strong correlation of more than 0.86 between the pairs indicating these libraries follow similar problem in all stages

API Misuses in all ML Stages

- API are often misused
- examined questions and accepted answers
- API misuses is observed in all stages of ML pipeline
- using wrong API
- Having API versions not match

RQ 3: Nature of libraries

- Eearly Stages for H2O and Mahout especially setup and model creation have comparetively higher percentage of questions compared to later stages

- Scikit-learn in an outlier in several categories suggesting that a deeper look into its API design might be necessary to improve usability of this important library

- Deep learning libraries caffe, H2O, keras, tensorflow, Theano, Torch show more training time difficulties compared to other ML libraries

RQ4 Time consistency of Difficulty

- Model creation related problems are consistent over time
- Data preperation related problems slowly decrease after 2013 and show sharp increase after 2017
- Training related problems shows slow increase over time
- Evaluation problems are consistent over time