ONNX Roadmap input

Andrew M. Sica, IBM Corp.

andrewsi@us.ibm.com

End to end pipeline support

Summary:

- Enable end to end pipeline using ONNX operators across both ONNX and ONNX-ML
- Continued focus on data preprocessing as part of ONNX / ONNX-ML and related convertors.
- Ability to combine multiple types (sklearn pipeline/tensorflow) enables a more embeddable / maintainable unit.

Thoughts:

- Opportunity to optimize data preparation without user rewriting.
  - Preprocessing is a pain point in enterprise use cases.
  - We've seen rewrites in golang, etc. in cases where latency (<10ms) matters.
- Continued work on coverage of pre-processing primitives and convertors.
- Multi-model support. can simplify deployment artifacts and governance.
- These can all be differentiators for the ONNX ecosystem.

Convertors

Summary:

- Focus on a single TensorFlow/Keras convertor (post tf2) is a positive step.
- Some loss of functionality around Keras LSTM, GRU layers, possibly others.
  - Decreased ability to optimize in ONNX backends for LSTM or GRU operations.
- Since the original input, we see that there are several related issues and indications that work is in progress:
  - https://github.com/onnx/tensorflow-onnx/issues/1684
  - https://github.com/onnx/tensorflow-onnx/issues/1546

Notes:

- We are investigating this further, beyond above examples.
- Example LSTM for fraud detection: https://github.com/IBM/ai-on-z-fraud-detection
- Our team will begin to provide additional feedback and look to contribute to TF-ONNX