ML Platform Evaluation Criteria

Data Science

Exported on 06/05/2020

Table of Contents

What is the functionality of a successful Machine Learning Platform?

1. What is the format of the output model?

1. Preferred format? PMML **M**
2. Support an API
3. Output model performance metrics **M**

- What metrics drive model performance? **M**

1. What are your options?

1. Have partial area under (AUC) the curve?

- Combine multiple metrics?
- Custom metrics (we choose)
- Platform modularity and extensibility

1. Can we add our own functionality to the platform? **M**
2. Can we invoke our own/external libraries from within the platform **M**
3. Can we run our own functionality hand-in-hand with platform functionality **M**
    a. The output of our functionality is accessible to the platform **M**
4. Does the platform have its own proprietary data store? Can we access files/info from outside the platform?

- Data Requirements **M**

1. structured **M**
2. unstructured data

1. Support for NLP frameworks, social media, graph analysis **M**
2. Video, images, speech

- Support for many data sets **M**

1. Train, validate, test, In-time, Out-of-time

- Is there a size limit for the input datasets?

1. i.e. is there a 100gb data file limit

- Will it handle our natural dataset sizes? **M**

1. 1 TB

- Will it handle big data? **M**

1. 4-5 TB

- Limitation on individual data sets/files (training, test, etc.) **M**
- Data streaming? **M**
- Software Requirements

1. Needs to run on and fully utilize Hadoop, Spark, Yarn, etc **M**

1. Needs to run on our environment specifically **M**
2. Pipeline execution in memory

- Support for Hive database (import ORC file type) **M**
- System Requirements **M**

1. OS - REL
2. Hadoop, spark, library dependencies?
3. Interfacing with a database
4. Interfacing with different data formats

- GDPR and PCI Compliance

1. Must not impede PCI compliance **M**
2. Mechanism to alert if there is PCI data in log files

1. Card numbers, bank account numbers, and PII in log file

- Cloud implementation **M**

1. Can it be implemented in the public cloud?
2. Can it be implemented in the private cloud?
3. Both public and private in tandem?

- Licensing

1. If the platform or any piece of it is installed at an ACI customer site

1. What are the licensing requirements/costs?

- Can it be licensed as a cloud based service to customers?
    a. From ACI cloud?
    b. From Public cloud?
- Licensing Model aspects
    a. ACI Internal Use (only ACI employees use it)
        i. Named user? Count based? Data size based? Etc.?
    b. ACI Customer Use (ACI customers use it)
        i. Named user? Count based? Tenant based? Data size based? Etc.?
- Pipeline Functionality **M**

1. Data Analysis and Visualization

1. Support for binary data
2. Aggregations on indices
3. Cross-tabs, frequencies, group by categorical variables, Chi2, Kramer's V, percentiles, mean, median, max, min
4. Box Plots, histograms, bar charts, heat maps, clustered bar charts, 3D plots

- Data Preparation, Pre-process, and Validation

1. Data Cleaning (missing value / imputation / outlier treatment)
2. Confirm user configuration of feature data (using user-defined logic)
3. Support for householding steps (as individuals), sorting
4. Joining and merging data sets
5. Filtering and Weighting data points
6. Derived variable calculation within the platform

1. create new features and new data fields
2. How difficult is it to calculate new features?
3. General Feature Generation / Engineering
    a. Automated feature engineering
    b. Manual feature engineering

- Labeling
- Identifying duplicates in the data
- Random Sampling - confirm the 3 steps of sampling

1. Can we ensure sample is representative

- Data manipulation (i.e cutting the fraud)
- Dataset management
    a. Keep track and update labels

- Dynamic Scaling

1. and batch scaling

- Feature Discretization
- Feature Range Optimization

1. Feature transformation

- Kernel Approximation
- Feature Selection

1. Support for Python or other FS libraries

- Modelling Algorithms

1. Support for multiple targeted ML libraries
    a. Python
    b. R **O**
    c. Scala **O**
    d. Spark
2. Automate algorithm selection

1. Manual algorithm selection

- User configurability of subset of algorithms to be automated
- User configurability of hyper-parameters for each algorithm to be automated
- Final Model Grooming

1. Take the best model for deep dive
2. Analysis across additional metrics
3. Analysis on additional data sets (OTT)
4. Analysis of model behavior

1. Analyze the incorrect model predictions

- Project Collaboration

1. Developer Environment

1. coder collaboration

- Notebook collaboration

1. which notebooks are supported?

- Track model performance

1. In production

- Model and Dataset Management

1. How are datasets and models organized?
2. Model archival?

- Model deployment

1. Automated
2. Manual
3. Use the model to score using any size dataset within the platform **M**

- Event logging **M**

1. Error handling

- Rules Engine?

1. Model constraints?