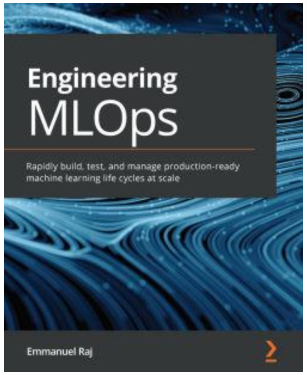
MLOps and Cloud Native AI/ML: Data and Machine learning operationalization

Presented by:

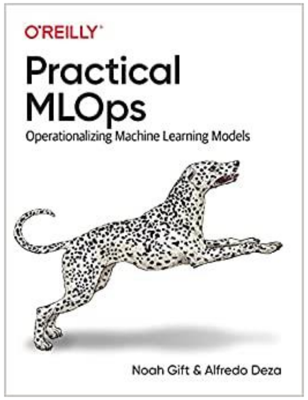
Prof. Fahd Kalloubi
Associate professor in data mining
and Big Data
Fahd.kalloubi@um6p.ma

- Associate professor in Data mining and Big Data at National school of applied sciences
- Ex Adjunct professor at euro-Mediterranean university of Fez
- Machine Learning and Big Data professional trainer
- Ex Professor at ENSIAS
- Research interests:
    - Data/Web mining and Natural language processing
    - Knowledge graphs and Machine learning/deep learning
    - Information retrieval and recommender systems

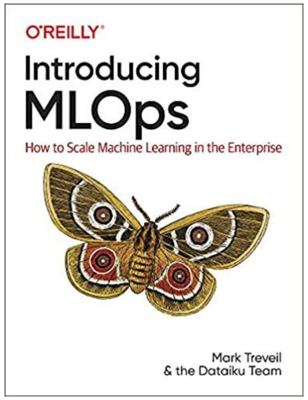
Literature



1. **Emmanuel Raj, Engineering MLOps. Packt publishing, 2021**

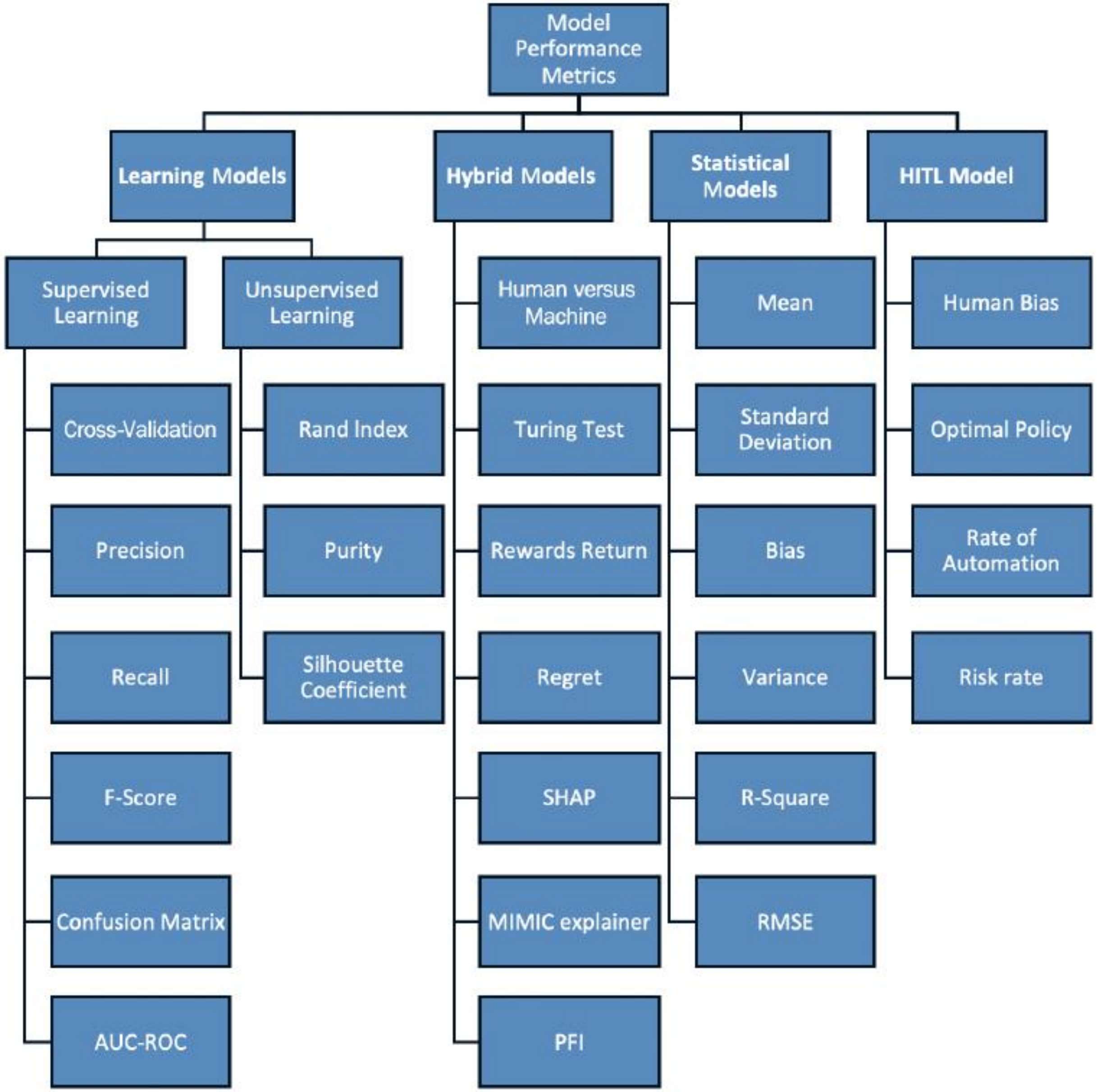


2. **Noah Gift and Alfredo Deza, Practical MLOps, O'Reilly publishing, 2021**



3. **Mark Treveil, and the Dataiku Team, Introducing MLOps, O'Reilly publishing, 2020**

Evaluation measures and methods for a ML model



Evaluating hybrid models

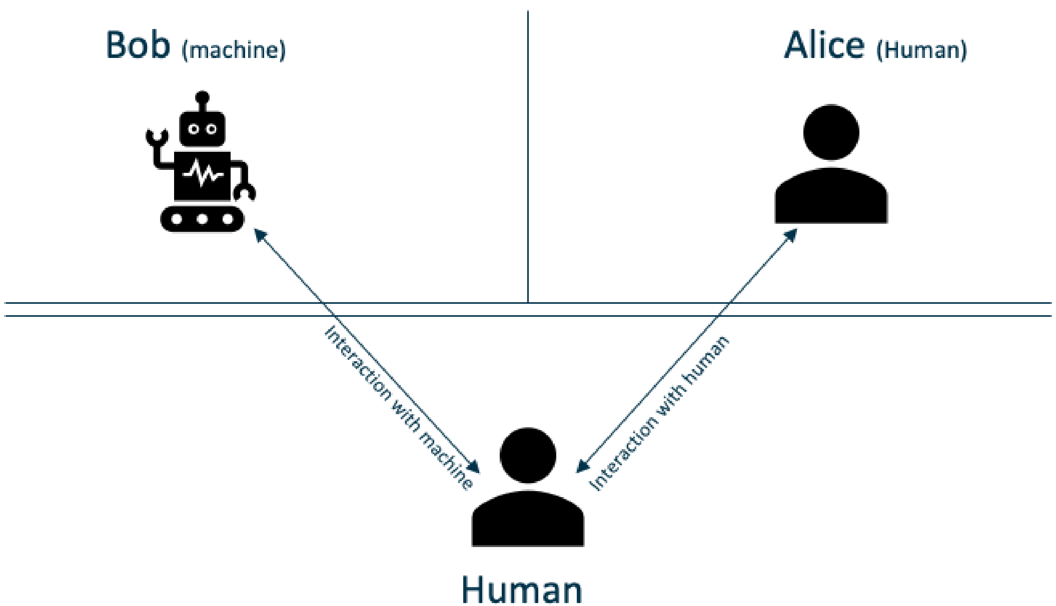
Human verus machine test

- These types of evaluations consist of comparing human performance against machine performance on a task.
- There are different metrics for evaluating human performance versus machines depending on context and tasks.
- Some examples:
  - **Bilingual evaluation understudy (BLEU):** is a method of assessing text quality for the task of machine translation from one language to another. The quality of text generated by a machine translation algorithm is compared to the output of a human.
    - The evaluation is carried out to observe how close a machine translation is to a professional human translation.
  - **Recall-Oriented Understudy for Gisting Evaluation (ROUGE) :** is a metric for evaluating human versus machine performance, used to evaluate tasks such as machine summarization and machine translation.
    - This metric compares a machine-generated summary or translation with a human-generated summary/translations

Evaluating hybrid models

Turing test

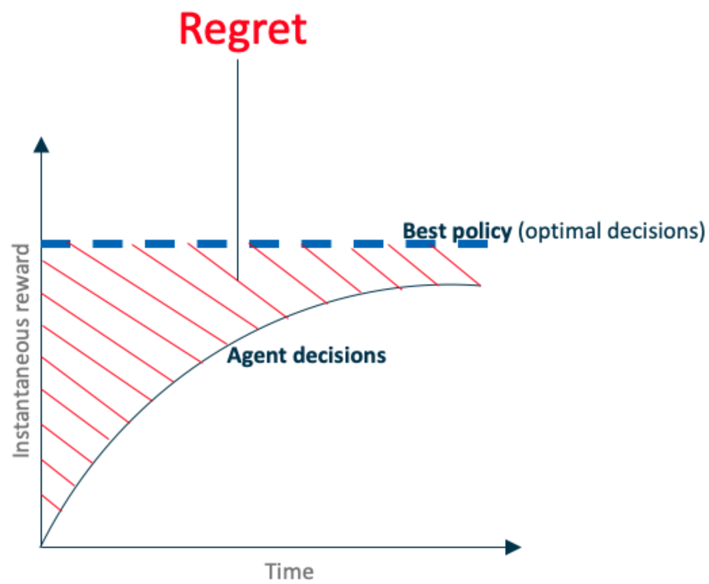
- The Turing Test is a test of a machine to assess its ability to exhibit intelligent, human-like behavior.
- It is also a test to evaluate the ability of a machine to deceive a human into believing that a task performed by a machine is human.



Evaluating hybrid models

Reward per return, Regret, SHAP

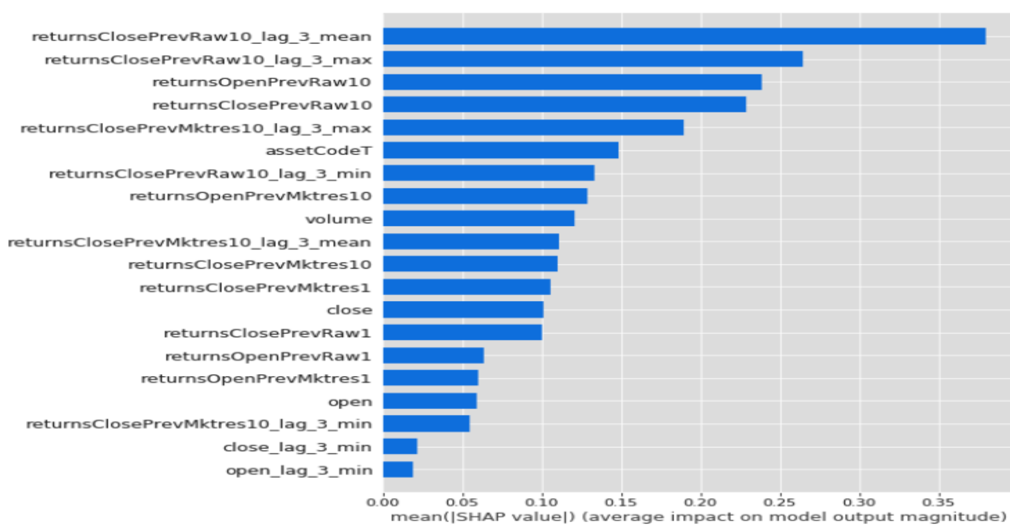
- Regret is a commonly used metric for hybrid models such as reinforcement learning models.
- At each time step, you calculate the difference between the reward of the optimal decision and the decision made by your algorithm. Cumulative regret is then calculated by summing.
- The minimum regret is 0 with the optimal policy. The smaller the regret, the better the performance of an algorithm.
- Regret allows us to evaluate the agent's actions in relation to the best policy



Evaluating hybrid models

Reward per return, Regret, SHAP

- Model interpretability and explaining why the model makes certain decisions or predictions can be vital in a number of problems
- Deep learning models are black box models.
  - We cannot explain their performance
- In such scenarios, the SHAP (**SHapley Additive exPlanations**) metric can be useful to decode what is happening with the predicted outcomes and which feature predictions are most correlated.
  - The main objective of SHAP is to explain the model output prediction by calculating the contribution of each feature
  - Output values describe the distribution of model outputs with respect to features



Evaluating hybrid models

Mimic explainer and PFI

- Mimic Explainer is an approach mimicking black box models by training an interpretable surrogate model.
- These trained surrogate models are interpretable models, which are trained to approximate the predictions of any black box model as accurately as possible.
- To train a surrogate model:
  1. Choose a dataset X, the same as the one on which the black box model was trained or another with a similar distribution
  2. Obtain the prediction of the black box model on the dataset
  3. Choose an interpretable model (linear model, decision trees, random forest, etc.)
  4. Using the dataset X and the predictions, train the interpretable model
  5. Evaluate how well the surrogate model reproduced the predictions of the black box model, for example, using R-squared or F-score.
  6. Obtain an understanding of the black box model predictions by interpreting the surrogate model.
- PFI (permutation feature importance) is an alternative to SHAP
  - Consists of randomly evaluating one characteristic at a time by calculating the change in the evaluation measures.
  - The change in the performance measure is assessed for each characteristic: the greater the change, the more important the characteristic.

Evaluation of HITL type models

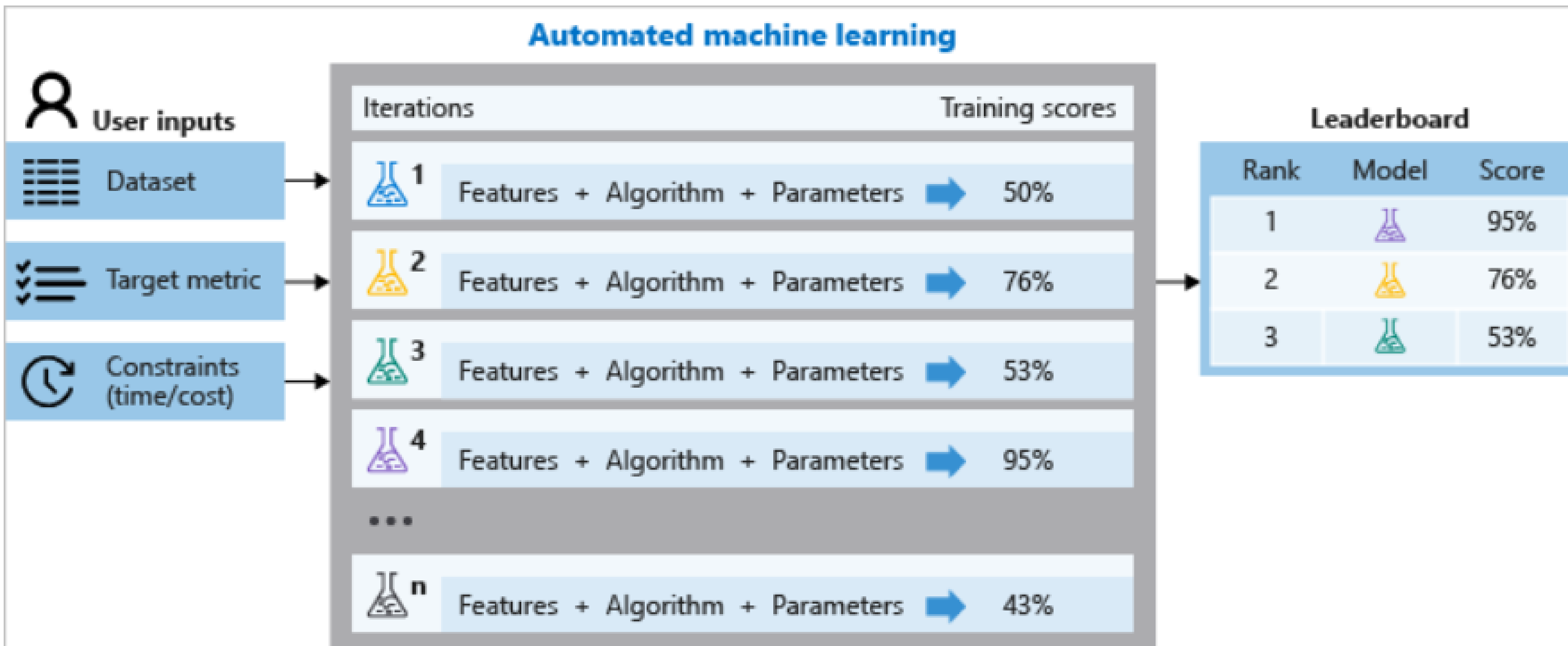
- Human biases
  - **Interaction bias**: When an ML system is fed a dataset containing entries of one particular type, an interaction bias is introduced that prevents the algorithm from recognizing any other types of entries
  - **Latent bias**: is experienced when multiple examples in the training set have a characteristic that stands out. Then, the ones without that characteristic fail to be recognized by the algorithm.
  - **Selection bias:** is introduced to an algorithm when the selection of data for analysis is not properly randomized
- There are other evaluation methods such as:
  - The optimal policy: In a system based on human reinforcement learning, a human operator or teacher sets the optimal policy, because the goal of the system is to achieve human-level performance.
  - Rate of automation: This is basically the percentage of tasks that are fully automated by the system (for example: DeepMind's AlphaGo achieved 100% automation to run on its own).
  - Risk rate: The goal of a human-in-the-loop HITL system is to reduce the error rate and teach the ML model to perform optimally.

Test methods in production

- Batch testing
  - Batch testing performed on a dataset to test model inference using metrics of choice, such as accuracy, RMSE, or f1-score.
  - In the cloud or on a remote server, the model is typically used as a serialized file and the file is loaded as an object for inference.
- A/B testing
  - When models are tested using A/B testing, the test will answer important questions such as:
    - ✓ Does the new Model B perform better in production than the current Model A?
    - ✓ Which of the two models works best in production to generate positive business indicators?
  - To evaluate the results of A/B tests, statistical techniques are used
- Stage test/shadow test
  - Before deploying a model for production, which would then lead to decisions being made, it may be interesting to replicate a production type environment (staging environment) to test the performance of the model.

Introduction to Azure AutoML

- AutoML (Automated machine learning), is the process of automating the iterative and time-consuming tasks of developing ML models.
- It enables data scientists, analysts, and developers to build ML models at scale, with high efficiency and productivity, while preserving model quality.
- AutoML processes in Azure:



https://docs.microsoft.com/en-us/azure/machine-learning/concept-automated-ml