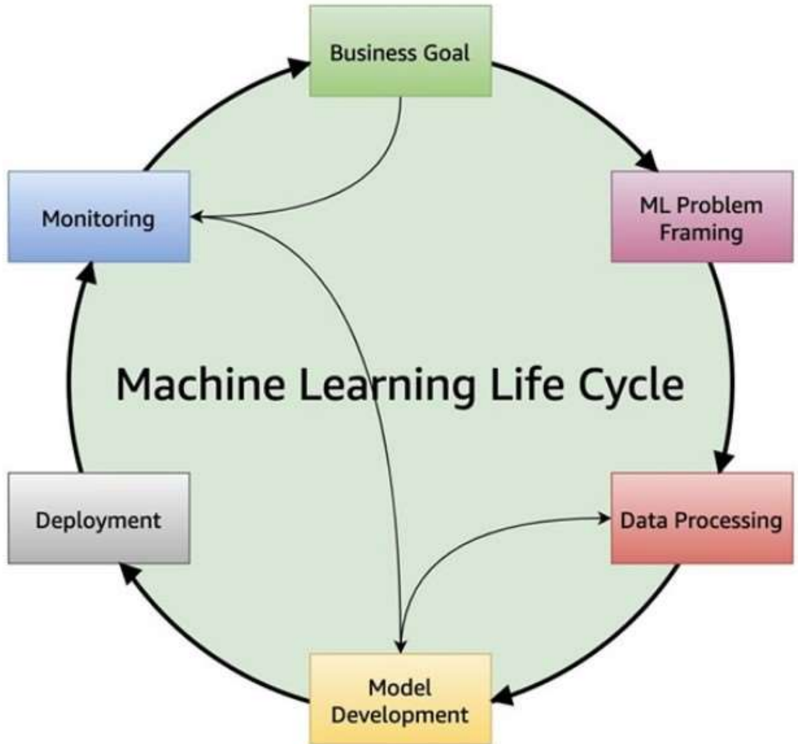
UVA | SCHOOL of DATA SCIENCE

ML Problem Framing

DS 6011: Capstone Part I / Capstone Prep
School of Data Science
University of Virginia

Last updated: September 8, 2023

ML Life Cycle



Source: AWS Well Architected Framework - ML Lens

ML Problem Framing

Following Business Goal step, we frame as ML problem

This assumes that ML is required for solution

In some cases, can solve with rules / analytics / queries

Goals of Problem Framing

1. Understand the use case and deliverables

2. Consider approaches (e.g., review literature)

3. Data collection / uses / availability. Predictions and target.

4. Performance metrics (e.g., # false positive for each true positive)

5. Responsible AI: building models that are safe and unbiased

Use Case and Deliverables

Clarify the use case: what needs to be done?

What does success look like?

What needs to be delivered? model / dashboard / etc.

Approaches

Which of these approaches is appropriate for the problem?

Type	Description	Labeling Requirement
Supervised learning	Learn a function to map predictors to target	Data needs to be labeled
Semi-supervised learning	Learn from labeled data, infer unlabeled data, and repeat	A portion of data needs to be labeled
Unsupervised learning	Learn groupings and outliers	Labels not required
Reinforcement learning	Learn from environment by taking action in a state, receiving next state and reward	Labels not required but reward function is needed

Literature

What papers may be helpful / relevant?

What code may be helpful / relevant?

Data

Is there documentation / data dictionary / data schema?

What data is available? Are there potentially useful predictors?

Is data labeled? Is there PII/PHI that needs to be masked?

How is data collected? Ideally, get walk-through of process.

What specifics should be known about data?

Understand how data is missing / limited / incorrect

Performance Metrics

Need to understand success criteria for adoption

Is there a known benchmark (e.g., current model in production with F1 score of 70%)?

Metrics will depend on type of problem (classification, regression)

Helps to think about impact to business for each type of error

1. False Positive (FP) – predicted positive but incorrect
2. False Negative (FN) – predicted negative but incorrect

Common Metrics – Binary Classifier

Metrics fall in range [0,1]. Higher is better.

Metric	Example
Recall fraction of positive cases detected	**Recall = #(true positive) / #(positive cases)** 100 customers that churn Of the 100, model predicted 70 as churn risk Recall = 70/100 = 0.70
Precision fraction of predicted positives that were correct	**Precision = #(true positive) / #(predicted positive)** 80 customers predicted to churn Of the 80 predicted, 70 churned Recall = 70/80 = 0.875
F1 score harmonic mean that balances recall, precision	**F1 = 2 (recall x precision) / (recall + precision)** = 2 (0.7 x 0.875) / (0.7 + 0.875) = 0.778

Common Metrics – Regression

R-squared

Fraction of variation in target variable explained by predictors.

Falls in range [0,1]. **Higher is better**

Adjusted R-squared

Fraction of variation in target variable explained by predictors, adjusted for model complexity.
Higher is better.

Root Mean Squared Error (RMSE)
Measures the average difference between predicted values and target values.
Lower is better. Sensitive to outliers.

Mean Absolute Error (MAE)
Measures errors between predicted values and target values.
Lower is better. Robust to outliers.

Responsible AI

ML brings automation at scale

A large risk is bias
Often results from under-representation of a group

Example: are there *protected classes* not represented?

Protected class: group of individuals protected by law
Examples: Age > 65, Female

Measuring Class Imbalance

Several ways to measure class imbalance

One simple measurement:
for given class, #(majority cases) - #(minority cases)

Another is based on outcomes for a group:
Example: Females 20% more likely to be rejected for mortgage

Mitigating Class Imbalance

Different approaches to mitigate class imbalance including:

1) collecting more data for minority class (not always possible)

2) resampling minority class with replacement (in training set)

References

AWS Responsible Use of Machine Learning

https://d1.awsstatic.com/responsible-machine-learning/responsible-use-of-machine-learning-guide.pdf

AWS Machine Learning Lens

https://docs.aws.amazon.com/wellarchitected/latest/machine-learning-lens/machine-learning-lens.html