Assessing Input Data and Resultant Model Accuracy

Saeid Molladavoudi (Lead Data Scientist)

Data Science Division, Statistics Canada

Government of Canada Data Conference

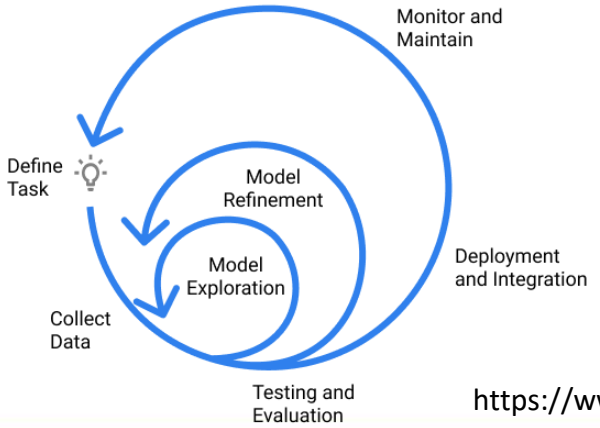
February 18, 2021

Delivering insight through data for a better Canada

General Considerations

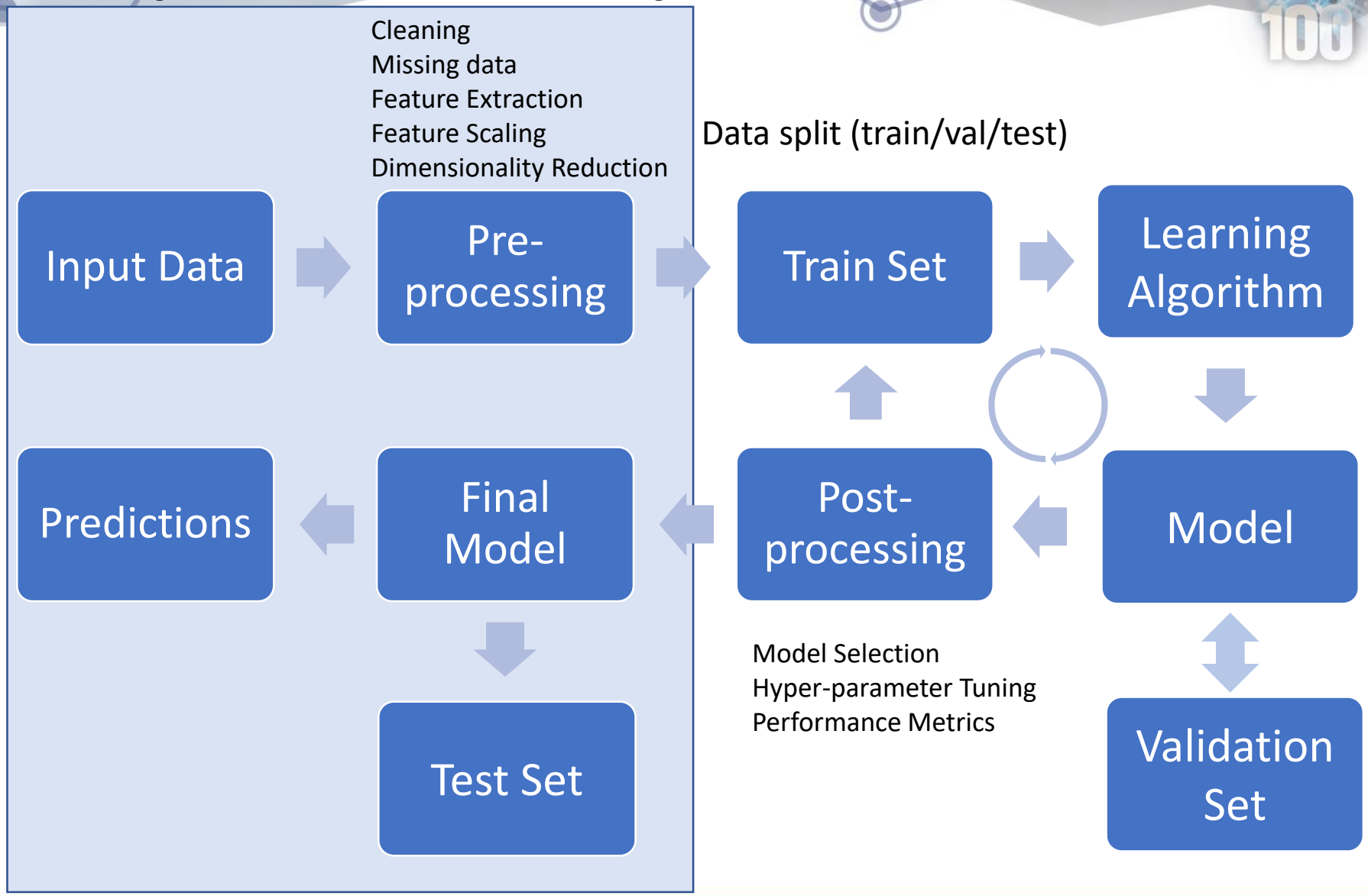
- **Every ML project starts with discussion with Subject Matters/data owners.**
- **The followings need to be clear from the beginning:**
  1. **Problem:** Identify the business need
  2. **Scope:** Supervised vs Unsupervised ML approaches
  3. **Objectives:** Classification, Regression, Clustering …
  4. **Constraints:** Concerns with the data (collection, amount, …)
  5. **Deliverables:** Machine-consumable vs Human-consumable solutions
  6. **Success criteria:** Manual hours or processing time saved, increases in quality, etc.

Machine Learning Development Lifecycle



https://www.jeremyjordan.me/ml-projects-guide/

Supervised Model Exploration Workflow



Input Data: most important part of a ML project

Main Challenges	• Not having enough data • Poor quality • Non-representative training data • Not enough relevant features • Too many irrelevant features
Questions that we need to answer	• Is there “enough” data? • How many sources of data exist? • How much data from each source? • Can more data be acquired? • Is data structured (e.g. via API) or unstructured (e.g. images)?
For instance, if the problem is a supervised ML use case	• Are there labels? • How much is the labeled data? • Can more labels be acquired?

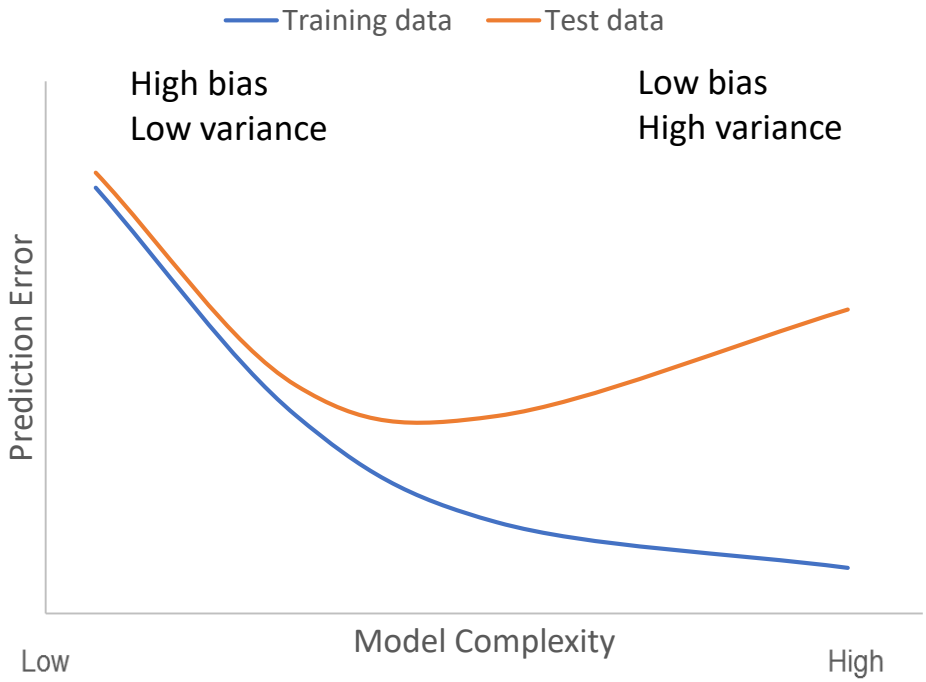
Exploratory Data Analysis (Pre-processing)

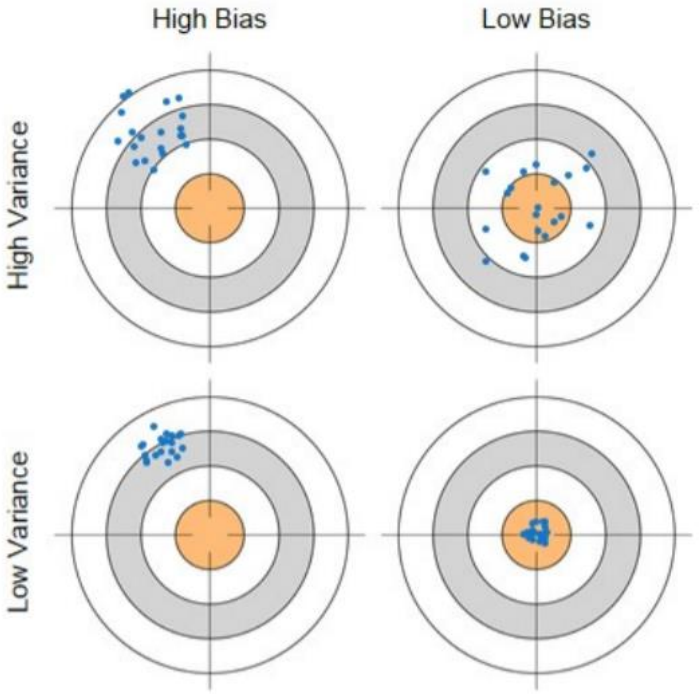
To consider	Details
Types of data	Heterogeneous (continuous, categorical, text, images, …)
Transformations	Feature engineering, categorical to one-hot, normalization, word embedding, doc-term matrix, …
Extraction	Pipeline if unstructured data, e.g. video, audio, pdfs, images …
Labels	Classification (binary, multi-class, multi-label) vs regression labels
Amount of data	Is there “enough” data to create train, validation and test sets?
Class Imbalance	Distribution of labels (class imbalance, changes over time)
Visuals	Visual tools, e.g. scatter plots and histograms
Summary statistics	Especially for structured data, e.g. to find the correlations among features
Missing values	Imputation or removal?
Dimensionality	Dimensionality reduction on the input data, e.g. Principle Component Analysis
Other types of tasks	Networks, time series, clustering, anomaly detection ….

Training, Validation and Testing Principle

Dataset	Purposes
Training Set	- The goal of ML is to minimize the expected loss (empirical risk minimization). - Models learn from training examples.
Validation Set	- Validation is used to find the best model parameters (hyper-parameter tuning).
Test Set	- Test is used to simulate what will happen when we apply the final model to new unseen data from the same distribution as the training data (i.e. real world). - It gives an unbiased estimate of generalization error of the model.

Bias-Variance Trade-off





https://www.variancejournal.org/articlespress/articles/Bias-Variance_Brady-Brockmeier.pdf

Supervised Classification ML

Data set size	Labels	Mitigation Strategies
Small	Few instances for minority classes.	• Use k-fold cross-validation. • Add regularization to the empirical risk to prevent over-fitting. • Use ensemble models. • Use data augmentation methods. • Transfer Learning. • Dimensionality reduction.
Medium	Few instances for minority classes.	• Use k-fold cross-validation. • Add less regularization. • Use ensemble models. • Use data augmentation methods. • Transfer Learning. • Learning curves to look for under vs overfitting. • Annotation tool.
Large	Small amount of labeled data.	• Semi-supervised learning. • Active learning. • Annotation tool.

Performance Evaluation

ML Task	Loss	Metrics
Classification	- (binary) cross entropy - Negative log-likelihood (log loss) - Hinge loss - Kullback-Leibler divergence - Relative entropy - Focal loss	- Accuracy - Precision - Recall - F-score - ROC curve - AUC
Regression	- Mean Squared Error (MSE) or quadratic loss - (Smooth) Mean Absolute Error (MAE) - Log cosh loss	- MSE - MAE
Clustering	- Within-cluster sum-of-squares	- Rand index - Mutual information score - Silhouette Coefficients
Ranking	- Margin ranking loss - Triple margin loss	- Mean reciprocal rank - Precision @ k - Normalized Discounted Cumulative Gain

Model Refinements:

- Use grid search or random sampling for hyper-parameter tuning.

Quality Assurance

Things to consider	Details
Uncertainty	• Aleatoric: Observation noise in data. • Epistemic: Uncertainty in model, parameters and convergence.
Bias in data	• Human bias in data labeling, e.g. in-group, homogeneity and implicit biases. • Collection bias that refers to the data that doesn't reflect the real distributions, e.g. convergence bias, reporting bias, participation bias and sampling bias.
Security and Privacy	• Privacy Preserving Technologies and their applications in ML, e.g. homomorphic encryption, secure multiparty computations, … • Input vs output privacy (confidentiality), e.g. differential privacy
Reproducibility	• Version control the code (development, testing, implementation) and data • Automated pipeline
Explainability	• Transparency • Documentation
Maintenance	• Monitor consistency, accuracy, coverage and representativity of learning data over time to detect any drift or slippage in quality

Thank you/Merci!

saeid.molladavoudi@Canada.ca