

# Recap – ML Fundamentals

Rina BUOY



AMERICAN UNIVERSITY  
OF PHNOM PENH  
STUDY LOCALLY. LIVE GLOBALLY.

# Statistics Mindset

The Data Scientist might have made assumptions about how the data is distributed, decided on a Linear Regression model, fitted the model, and diagnosed the model.

Now the Business Analyst may interpret the parameter estimates, including confidence intervals and hypothesis tests.

This model is optimized to infer the original distribution of the data and all metrics that comes with it.

The fact that you apply this model to new data without thoroughly examining this data is something that scares every frequentist statistician.

# Machine Learning Mindsets

In Machine Learning, the regression model came out of a “contest” between multiple models.

The Linear Regression model just happened to be the most performative one.

Performance is usually measured by comparing the performance of a model on training and testing datasets.

The application and evaluation of the model on new, real data is actually nothing that ML practitioners fear but something they embrace, as that’s one of the main goals.

Machine Learning heavily relies on out-of-sample metrics to evaluate the performance of a Machine Learning model.



The Analyst could use the model to predict new data points, but interpreting the parameter estimates or calculating confidence intervals might not be a good idea.

# Statical Evaluation

## Training Data

- Used to fit models (estimate parameters).
- Typically, 60-80% of the data set.

## Validation Data

- Used to measure the **error of candidate models**.
- Training and validation are performed iteratively until model achieves the desired performance.

## Test Data

- Used to measure the **performance of the final selected model**.
- True blind test as validation data was used multiple times during training.

# Dealing with Imbalanced Dataset - Target Variable Strategies

Oversampling the minority class

- Randomly duplicate samples from the minority class until the classes are balanced out.

Undersampling the majority class

- Remove samples from the majority class randomly until both classes contain an equal number of observations.

Generate synthetic data

- Create new, synthetic samples for the minority class by interpolating existing data points.

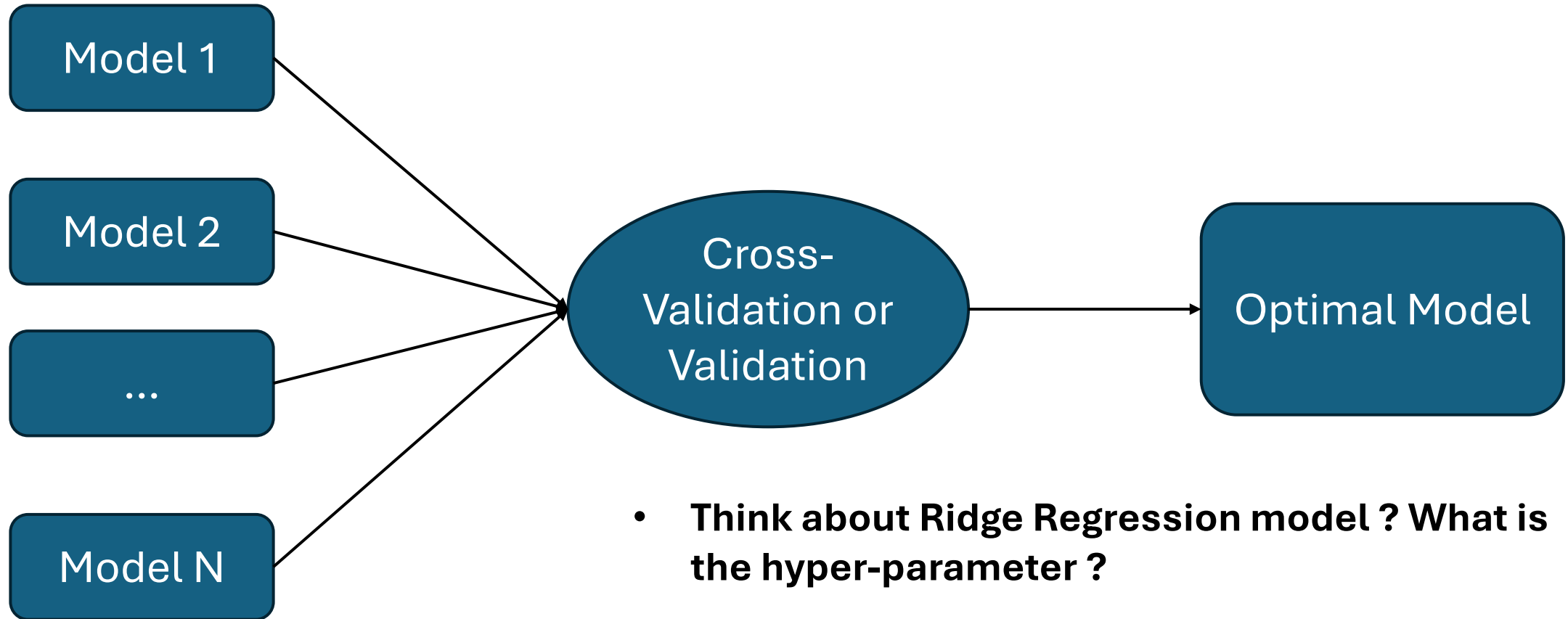
Penalize algorithms

- Modify learning algorithm to consider class imbalance.

Try a different algorithm

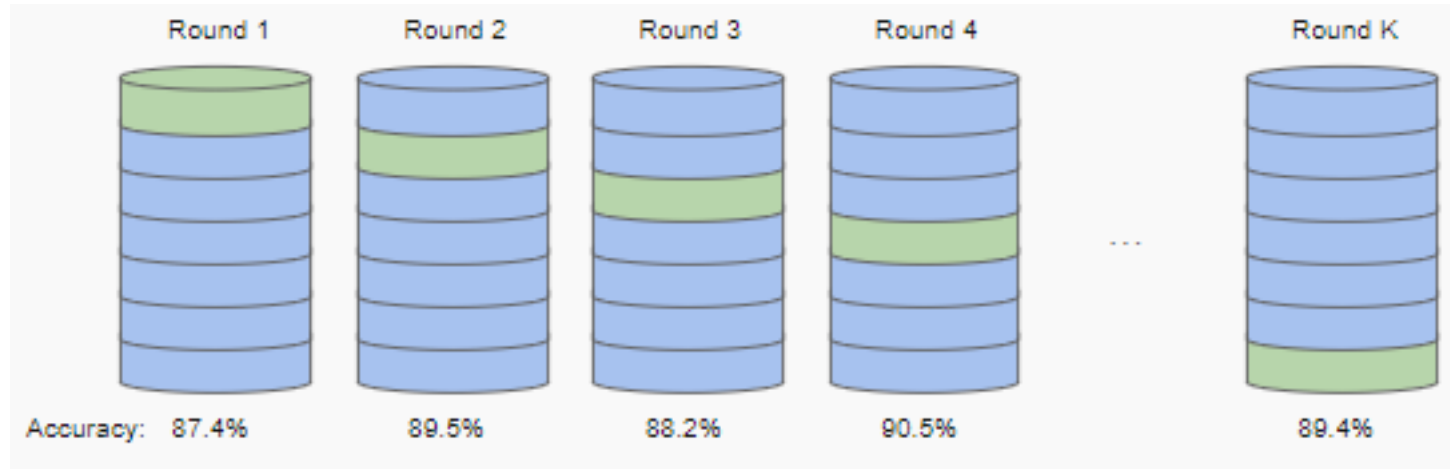
- Algorithms that usually work well on imbalanced data: random forests, boosted trees, naive bayes classifiers, K nearest neighbors.
- Algorithms that usually do not work well on imbalanced data: linear regression, logistic regression, neural networks.

# Model Selection



- **Think about Ridge Regression model ? What is the hyper-parameter ?**
- **Think about Polynomial Regression model ? What is the hyper-parameter ?**

# Cross Validation – For Model Selection/Hyper-parameter Tuning



Split the data into K folds.

K-1 folds are used for training, 1-fold for validation.

Shuffle the validation fold k times.

# Alternatively





