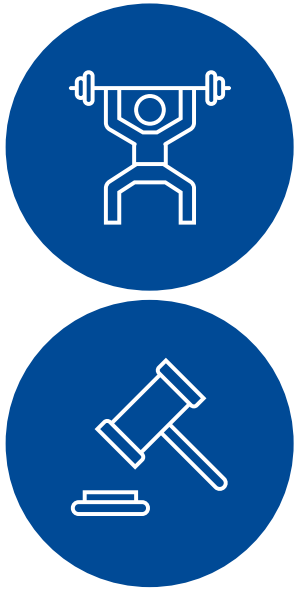
An Introduction to Artificial Intelligence (AI) in Finance

Chapter 3:
Training, Validation, and Testing

Universität
Münster

3. Training, Validation, and Testing



1. Model Performance Measures
2. Generalization & Overfitting
3. Train-Test-Validation-Split
4. Hyperparameter Optimization
5. Assessing Real-World Impact

3.1 Model Performance Measures

Definition

Model Performance: Refers to how well a machine learning model makes predictions or decisions based on data.

> **!** A good choice of model performance measure considers its effectiveness in measuring the quality of the prediction, its robustness to noise or outliers, its computational efficiency, and its interpretability.

3.1 Model Performance Measures
Common Classification Metrics

$$Accuracy = \frac{TP + FP}{TP + TN + FP + FN}$$

Proportion of correct predictions.

$$Precision = \frac{TP}{TP + FP}$$

Proportion of positive predictions that were actually correct.

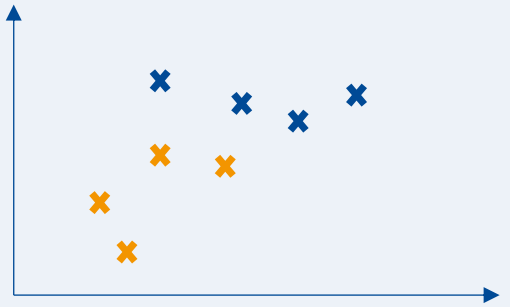
$$Recall = \frac{TP}{TP + FN}$$

Proportion of actual positives that were correctly predicted.

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Harmonic mean of precision and recall.

Classification:
Predicting categorical outcomes (e.g., credit risk classification).



− Struggle with imbalanced datasets.

\+ Relatively easy to understand

3.1 Model Performance Measures
Common Classification Metrics

Area under the Curve (AUC): Quantifies the overall ability of the model to discriminate between positive and negative classes.

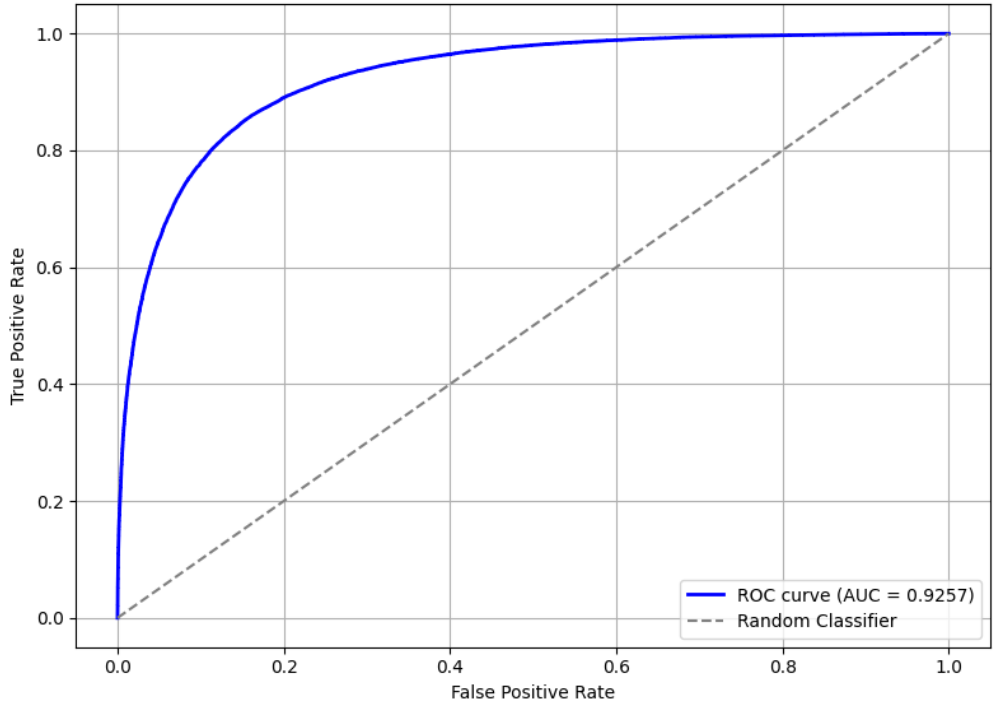
Trade-off between TPR and FPR

True Positive Rate

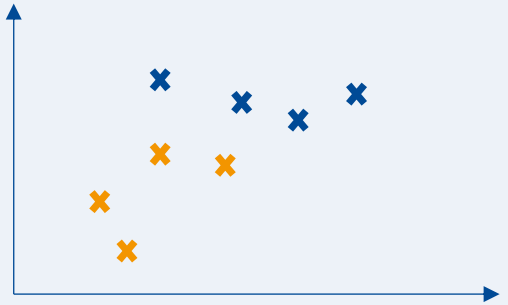
$$TPR = \frac{TP}{TP + FN}$$

False Positive Rate

$$FPR = \frac{FP}{TP + FN}$$



Classification:
Predicting categorical outcomes (e.g., credit risk classification).



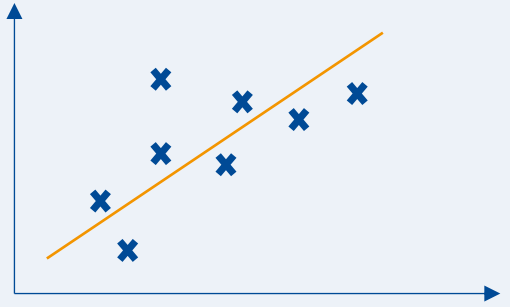
− Less intuitive than other metrics

\+ Useful for imbalanced datasets

3.1 Model Performance Measures

Common Regression Metrics

Regression:
Predicting continuous outcomes (e.g., stock price prediction).



Mean Squared Error

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- \+ Penalizes large errors
- − Sensitive to outliers

Root Squared Error

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- \+ Penalizes large errors, same unit as the target
- − Sensitive to outliers

Mean Absolute Error

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- \+ Robust to outliers, same unit as the target
- − Harder to optimize

Coefficient of Determination

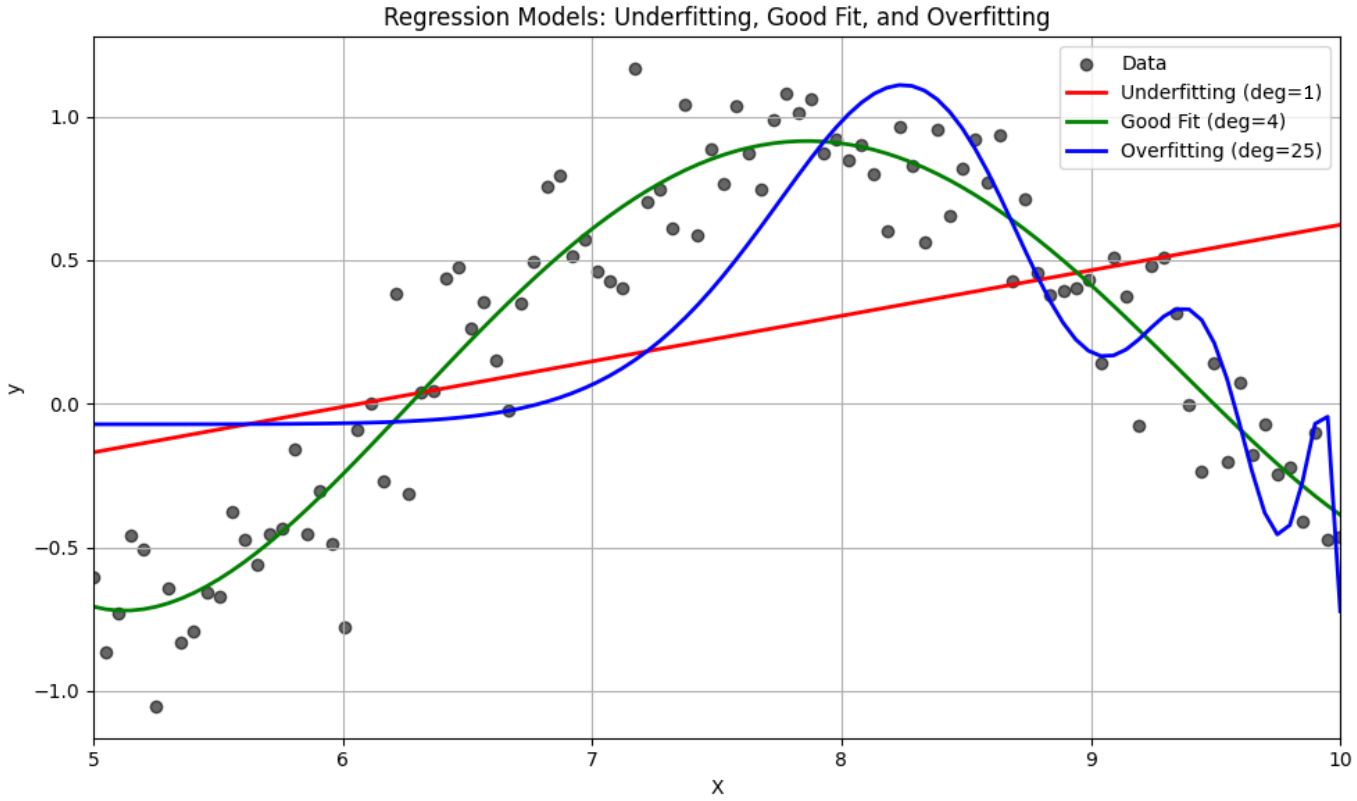
$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- \+ Normalizes, comparative within dataset
- − Potentially Misleading

3.2 Generalization and Overfitting

Generalization: A model's ability to perform well on new, unseen data by capturing the underlying patterns within the data.

Overfitting: Situation in which a model fits the training data (including noise and outliers) too close so that its ability to generalize deteriorates.



! Potential remedies for overfitting include using more training data, reducing model complexity, using regularization techniques, and applying cross-validation.

3.3 Train-Validation-Test-Split
Sampling Methods

Train-Validation-Test-Split: The process of dividing a dataset into subsets for training, validating, and testing machine learning models to minimize overfitting and ensure that the model generalizes well to unseen data.

Training Data
Subset used to train the model

Validation Data
Subset used for model selection and to fine-tune hyperparameters

Testing Data
Subset used for an unbiased performance estimation of the final model on unseen data

Simple Random Sampling: Randomly dividing the dataset into training, validation, and testing subsets.

Stratified Sampling: Dividing the data in such a way that the proportion of classes in the training, validation, and testing sets reflects the overall class distribution.

3.3 Train-Validation-Test-Split

Holdout Validation

Holdout Validation: Method for splitting data where data is simply divided into three different sets and assigned to training, validation, and testing.

- ⊕ Simple and efficient
- ⊖ High variance of results

3.3 Train-Validation-Test-Split
K-Fold Cross-Validation

K-Fold Cross-Validation: Method for splitting up data by dividing the dataset into 'k' equal-sized folds:

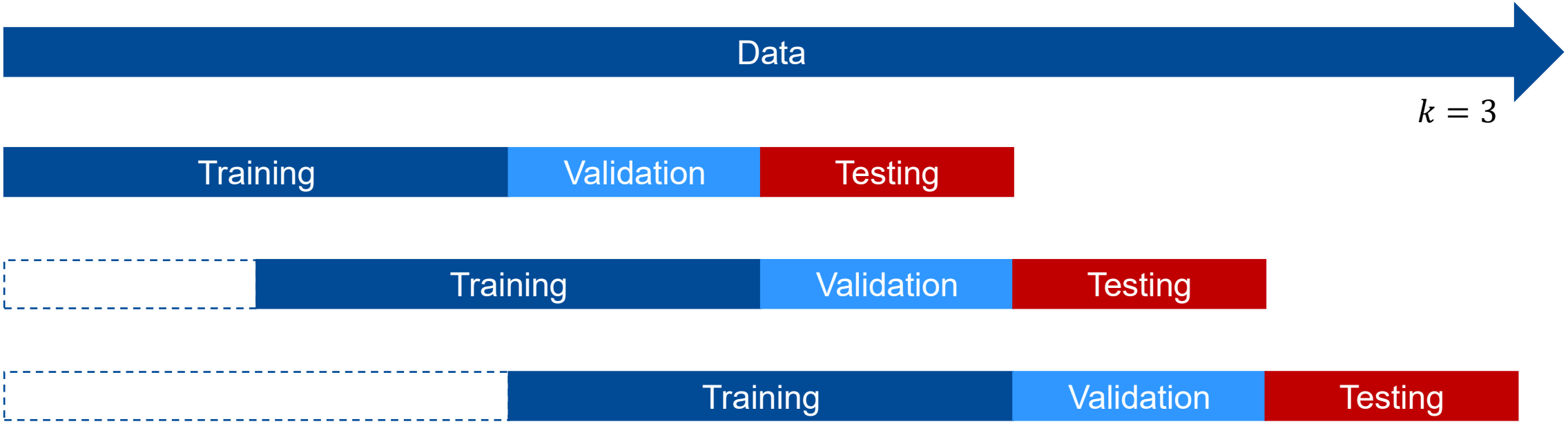
1. Separate the test data
2. Divide the remaining dataset into 'k' equally sized folds.
3. For each fold, use 'k-1' folds for training and the remaining fold for validation.
4. Repeat this process 'k' times, ensuring that each fold serves as the validation set once.
5. Calculate the average validation performance over all 'k' iterations.

3.3 Train-Validation-Test-Split
K-Fold Cross-Validation

- ⊕ Uses all data points both for training and validation; more comprehensive measure for model performance.
- ⊖ Computationally intensive and time consuming.

3.3 Train-Validation-Test-Split

Cross-Validation for Time-Series Data



3.4 Hyperparameter Optimization
Definitions

Model-Parameter: Parameter that is automatically tuned during model optimization.

(e.g., coefficients in a linear regression, weights in a neural network)

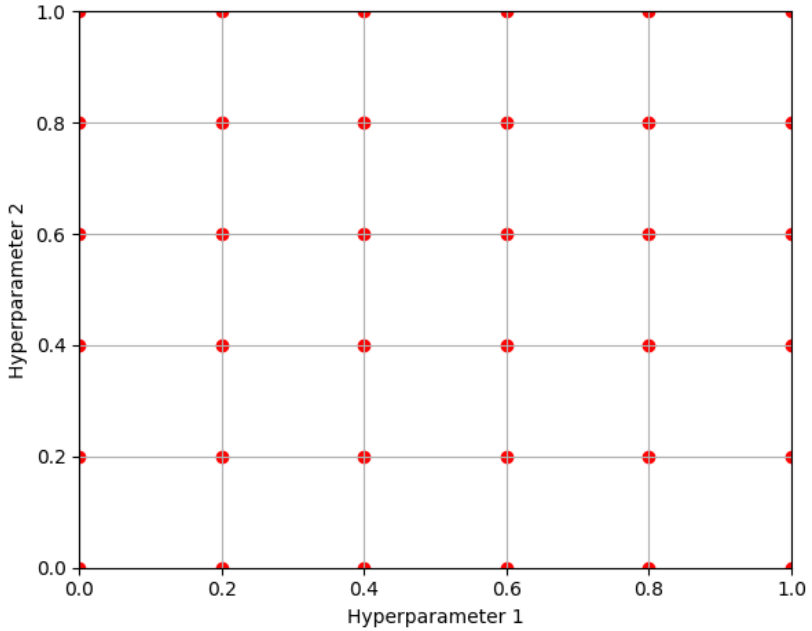
Hyperparameter: Configuration value that is set before optimizing a machine learning model.

(e.g., learning rate of the optimizer, maximum depth of a decision tree)

Hyperparameter Optimization: Process of finding the best combination of hyperparameters that leads to the best model performance on the validation set.

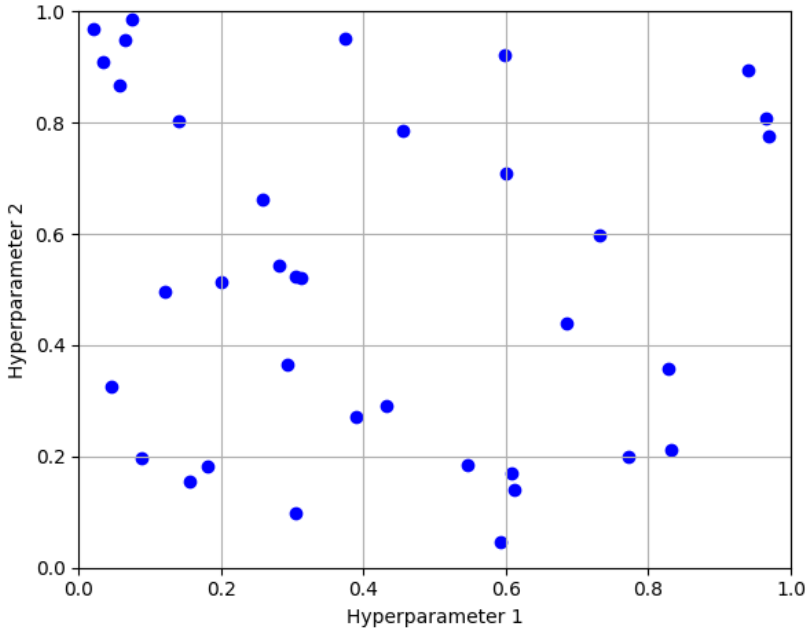
3.4 Hyperparameter Optimization

Approaches



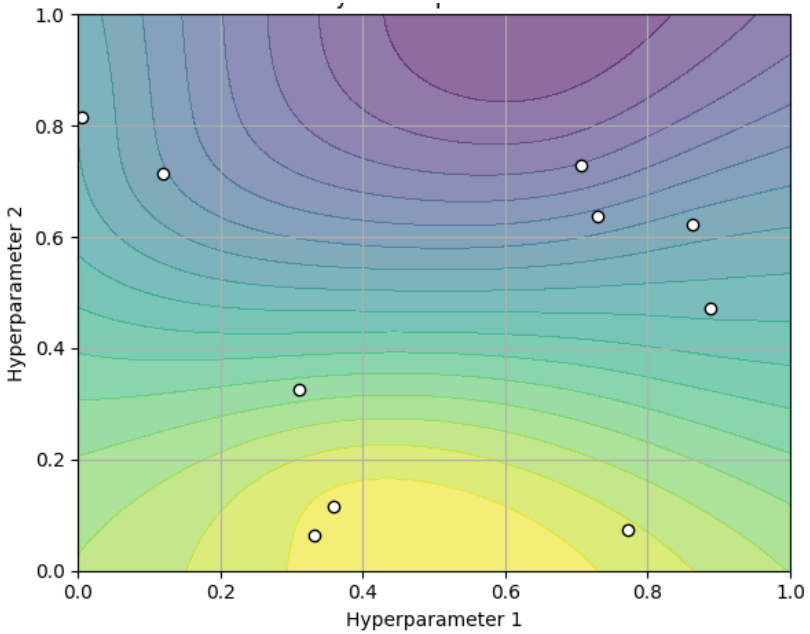
Grid Search: Samples points in a uniform grid across the space.

- ⊕ Exhaustive
- ⊖ Inefficient



Random Search: Samples points randomly.

- ⊕ More efficient

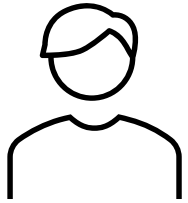


Bayesian Optimization: Uses a probabilistic model to guide sampling toward promising regions.

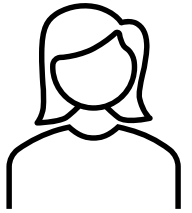
- ⊕ Very efficient
- ⊖ Struggles in high dimensions

3.5 Assessing Real-World Impact

“I built a churn prediction model with 95% accuracy on test data.”



“Does this model reveal new patterns in customer behavior that weren’t previously known?”



“If we act on this model’s predictions, how many customers will we retain and at what cost?”

E3 – Classification Metrics

Below is a confusion matrix of test data for a logistic regression model predicting companies' defaults (default = 1, no default = 0). Calculate accuracy, precision, recall, and F1 score. Argue which metric would be most relevant if missing a default is more important than a false alarm. In this case, should you only focus on this individual metric?

		Prediction	
		Default	No Default
Reality	Default	80	5
	No Default	10	105