Project documentation and reporting. Formal and modelling languages for AI project documentation

Typer of Documentation in AI

- Process documentation
  - According to steps in project lifecycle:
    - Goal identification -> Data colelction and preparation->Feature engineering... (Lecture 2)
- Technical documentation
  - Methodologies, algorithms, models, tools, techniques;
- User documentation
  - End-user documentation (interfaces, instructions, system behaviour)
- Reports
  - Findings, progress, performance metrics, key deliverables;

Modeling Languages for AI Projects

- UML (Unified Modelling Language)
- BPMN (Business Process Model and Notation)
- SysML (Systems Modelling Language)
- Domain-Specific Modelling Languages (DSLs)

Process documentation

- https://bit.ai/templates/process-documentation-template
- https://github.com/eugeneyan/ml-design-docs

Technical documentation

- System Architecture Diagram
  - Overview of an entire AI system
  - Show how different components interact
  - Data sources, data prepocessing and storage components, model training pipelines(), deployment architecture (cloud services, serverless functions, edge devices)
  - https://www.restack.io/p/ai-frameworks-answer-ai-architecture-diagram-cat-ai
  - Papers with diagrams:
    - https://www.researchgate.net/publication/267953355_Eagle_Knights_2006_Small-Size_League
    - https://www.researchgate.net/publication/323405510_Convergent_Innovation_in_Food_through_Big_Data_and_Artificial_Intelligence_for_Societal-Scale_Inclusive_Growth

Technical documentation

- Data Flow Diagram (DFD)
  - Designed to show how data flows within the system
- Entity-Relationship Diagram (ERD)
  - How entities are related to each other
- Sequence diagrams
  - https://zenuml.com/
- Activity Diagram
  - Verious activities in the AI pipeline and decision points
- Model architecture diagram
  - Layers, connections between layers, hyperparameters and activation functions
- Pipeline diagram

DAG (directed acyclic graph)

- DAG - a mathematical abstraction of a pipeline
- Tools:
  - Apache Airflow
  - Argo (https://argoproj.github.io/)
  - Prefect (https://www.prefect.io/)
  - Luigi (https://luigi.readthedocs.io/en/stable/index.html)

Technical documentation

- Component diagram
- Confusion matrix (Performance evaluation)
  - https://scikit-learn.org/stable/
- State diagram
- Deployment diagram
  - Shows how the AI system is deployed in a production environment, including servers, databases, cloud infrastructure, etc.

Reports

- Findings examples:
  - Model performance: The model trained on a dataset of 50,000 labeled images for object detection achieved an accuracy of 92%, which is higher than the baseline model's accuracy of 85%. However, the model shows reduced performance (89%) on images taken under low-light conditions
  - Feature Importance:Finding: The most important feature for predicting house prices is the "square footage" of the house, followed by the "neighborhood rating" and "number of bedrooms." Other features like "yard size" and "distance to city center" had minimal effect on predictions.

Reports

- Performance Metrics
  - Example – Classification model:
    - Accuracy: 92%
    - Precission: 0.89
    - Recall: 0.91
    - F1-score, AUC, Confusion matrix
- Key deliverables - important milestones, outputs, and tangible results of the project

Practice: Management and documentation of evaluation metrics and performance assessment

- Model Evaluation

Title	Details
Model name	Name…
Algorithm	Random forest, NN, …
Training data size	Number of training samples
Test data size	Number of test samples
Date of evaluation	Date…
Evaluation method	Cross-validation…

Classification Model Evaluation

- Metrics for Binary and Multi-class Classification: Accuracy, recall, precision, F1-score, specificity, ROC-AUC, PR-AUC, Log Loss
- Metrics for Imbalanced Datasets: Balanced Accuracy, Cohen's Kappa, Matthews Correlation Coefficient (MCC), F-beta Score

- Visuals: confusion matrix, ROC curve, precision-recall curve

Regression Model Evaluation

- Error-based Metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), Median Absolute Error
- Goodness-of-Fit Metrics: R-squared (Coefficient of Determination), Adjusted R-squared
- Residual Analysis: Histogram or Q-Q plot of residuals, Residual vs. Predicted Value plot

Clustering Model Evaluation

- Internal Evaluation Metrics: Silhouette Score, Dunn Index, Calinski-Harabasz Index
- External Evaluation Metrics: Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), Fowlkes-Mallows Index

Time Series Model Evaluation

- Error Metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), Symmetric Mean Absolute Percentage Error (SMAPE), Mean Absolute Scaled Error (MASE)
- Specialized Methods: Cross-Validation with Rolling Window, Diebold-Mariano Test

Recommender System Evaluation

- Ranking-based Metrics: Mean Reciprocal Rank (MRR), Normalized Discounted Cumulative Gain (NDCG), Hit Rate
- Prediction Accuracy Metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE)
- Coverage metrics

General Cross-Validation Methods

- K-Fold Cross-Validation
- Stratified K-Fold
- Leave-One-Out Cross-Validation (LOOCV)
- Repeated K-Fold
- Time Series Split
- Bootstrap Sampling

Model Interpretability and Fairness Evaluation

- SHAP
- LIME
- Fairness Metric

Practice: Management and documentation of evaluation metrics and performance assessment

- Model Evaluation continued:
  - Metrics used:
    - Name of metric
    - Formula
    - Result
    - Interpretation
  - Confusion matrix

Practice: Management and documentation of evaluation metrics and performance assessment

- Performance assessment
  - Model overview

Title	Description
Model name	
Model type	
Problem type	
Model version	

Practice: Management and documentation of evaluation metrics and performance assessment

- Performance assessment
  - Quantitative metrics

Title	Description
Accuracy	
Precision	
Recall	
AUC-ROC	
Etc.	

  - Inference time (Average Inference Time per Sample, Throughput (Samples per Second), Use Case Impact)

Practice: Management and documentation of evaluation metrics and performance assessment

- Performance assessment
  - Training time - Total Training Time, Training Speed, Scalability Assessment
  - Memory usage - Memory Consumption. Model Size, Optimization Techniques Used
  - Model Robustness and Reliability - Stress Testing, Model Drift Detection, Error Analysis
  - Deployment Readiness