ML Workflow and Best Practices

ML Workflow & Problem Definition

Learning Objectives

By the end of this lesson, you will be able to:

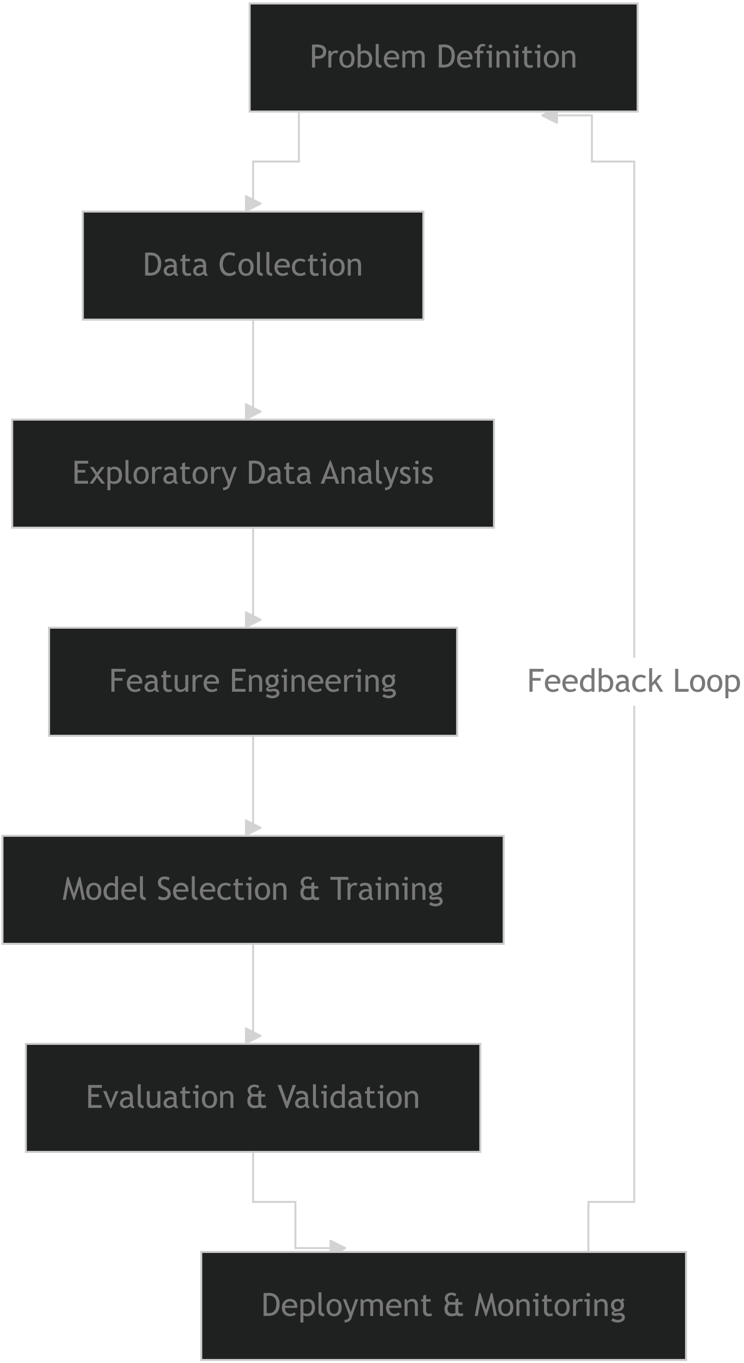
- Define the machine learning (ML) workflow and its key stages.
- Identify the essential steps for framing an ML problem.
- Align ML problem definitions with specific business objectives.
- Recognize common challenges in ML projects and how to address them.

The Machine Learning Workflow

A typical ML workflow consists of the following stages:

- **Problem Definition & Business Understanding** - Define the goal, constraints, and success criteria.
- **Data Collection & Preprocessing** - Gather and clean data for modeling.

- **Exploratory Data Analysis (EDA)** - Understand the dataset through statistics and visualizations.
- **Feature Engineering** - Select and transform relevant features for better model performance.
- **Model Selection & Training** - Choose and train an appropriate model.
- **Evaluation & Validation** - Measure model performance using appropriate metrics.
- **Deployment & Monitoring** - Deploy and continuously monitor model performance.

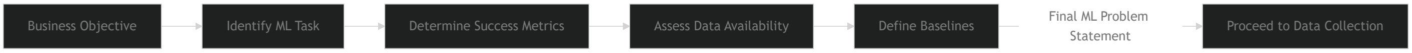


Framing an ML Problem

The success of an ML project hinges on how well the problem is defined. Poorly framed problems often lead to ineffective solutions.

Key Steps to Define an ML Problem:

1. **Understand Business Objectives** - What is the end goal? (e.g., increasing revenue, reducing churn, automating tasks).
2. **Identify the ML Task** - Is it a classification, regression, clustering, or recommendation problem?
3. **Determine Success Metrics** - What defines a “good” model? (e.g., accuracy, precision-recall, RMSE, F1-score).
4. **Assess Data Availability & Constraints** - Do we have enough data? Is it labeled? Are there ethical considerations?
5. **Define Baselines** - What is the current performance without ML? (e.g., rule-based systems, human decision-making).


Business Objective
Identify ML Task
Determine Success Metrics
Assess Data Availability
Define Baselines
Final ML Problem Statement
Proceed to Data Collection

3Common Challenges in ML Projects

- **Data Issues:** Missing, biased, or insufficient data.
- **Feature Engineering Complexity:** Selecting the right features impacts model performance significantly.
- **Model Generalization:** Avoiding overfitting or underfitting

Defining an ML Problem with Data

Using Python, explore a small dataset and determine an appropriate ML problem statement.

Steps:

1. Load a dataset (e.g., customer purchase history, loan applications).
2. Analyze key features using Pandas.
3. Determine if the problem is classification, regression, or clustering.
4. Print a structured ML problem statement.

Example Code:

Copy

```
import pandas as pd

# Load dataset
url = "https://raw.githubusercontent.com/mwaskom/seaborn-data/master/titanic.csv"
df = pd.read_csv(url)

# Display dataset info
df.info()
print("\nSample Data:\n", df.head())

# Identify ML problem type
if 'survived' in df.columns:
    print("Potential ML Problem: Binary Classification (Predicting survival)")
```

Key Takeaways

- A well-defined ML problem is crucial for project success.
- Aligning ML solutions with business objectives ensures practical impact.
- Understanding data constraints early helps mitigate risks.
- Framing the problem correctly leads to better model performance and deployment outcomes.

< Previous

Attributions

Next >