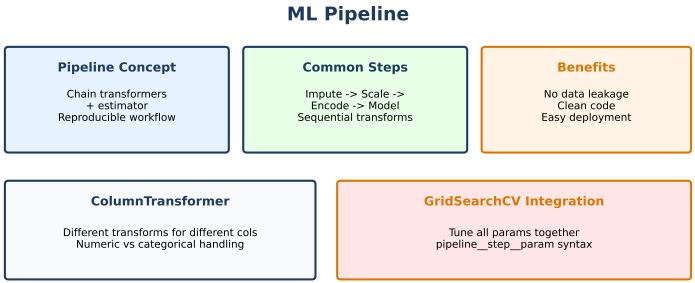
Lesson 32 Summary: ML Pipeline

Data Science with Python – Key Concepts

Data Science Program

ML Pipeline Overview



Pipeline([("scaler", StandardScaler()), ("model", LogisticRegression())])

Pipelines chain preprocessing and modeling steps

Why Pipelines?

Benefits:

- **No data leakage:** Proper train/test separation
- **Reproducibility:** Single object captures workflow
- **Deployment:** Easy to save and load

Pipelines prevent common ML mistakes

Pipeline Structure

Chain of steps:

- **Transformers:** Scale, encode, impute
- **Estimator:** Final model (must be last)
- **Sequence:** Each step transforms data for next

All steps except last must have transform method

ColumnTransformer

Different transforms per column type:

- **Numeric:** Scale, impute with mean
- **Categorical:** One-hot encode, impute with mode
- **Parallel:** Apply transforms simultaneously

ColumnTransformer handles mixed data types

Common Pipeline Pattern

Typical workflow:

- **Step 1:** Handle missing values
- **Step 2:** Scale numeric features
- **Step 3:** Encode categoricals
- **Step 4:** Fit model

This pattern covers most ML preprocessing needs

GridSearch with Pipelines

Tune all parameters together:

- **Syntax:** step__parameter
- **Example:** model__C for regularization
- **Benefit:** Finds optimal preprocessing + model

Double underscore separates step name and parameter

Saving and Loading

Deployment-ready:

- **joblib:** Save entire pipeline
- **Load:** Full preprocessing + model
- **Predict:** Works on raw data

Saved pipeline includes all fitted transformers

Best Practices

Pipeline tips:

- **Name steps:** Use descriptive names
- **Test early:** Verify with small data
- **Document:** Record preprocessing choices

Clear naming makes debugging easier

Quick Reference

Essential Commands:

Task	Code
Create pipeline	`Pipeline([('scaler', ...), ('model', ...)])`
Fit	`pipe.fit(X_train, y_train)`
Predict	`pipe.predict(X_test)`
Grid search	`GridSearchCV(pipe, param_grid)`
Save	`joblib.dump(pipe, 'model.pkl')`

Pipelines are essential for production ML