01_pipeline_demo

June 30, 2025

Cells and import

[2]:

```python
from pathlib import Path

from credit_risk.data_ingest import load_data, clean_data
from credit_risk.features import generate_features

# notebook cwd is .../credit-risk-project/notebooks
notebook_path = Path.cwd()

# project_root is one level up
project_root = notebook_path.parent

raw_csv = project_root / "data" / "raw" / "borrowers.csv"

df_raw = load_data(raw_csv)
df_clean = clean_data(df_raw)
```

```
Loaded 5 rows and 6 columns from /Users/glennasher/credit-risk-
project/data/raw/borrowers.csv
Dropped 0 duplicate rows
Filled missing 'employment_length' with median=7.0
Filled missing 'age' with median=45.0
Filled missing 'annual_income' with median=60000.0
Clipped 'debt_to_income' to [0,1]
One-hot encoded columns: ['emp_bin', 'age_bin']
Cleaned data has 5 rows and 16 columns
```

Feature generation

[3]:

```python
x_train, x_test, y_test, y_train = generate_features(
    df_clean,
    target_col = 'default',
    test_size = 0.4,
    random_state = 42,
)
```

```
Feature generation complete: 3 training samples, 2 test samples
Scaled numeric columns: ['annual_income', 'employment_length', 'credit_score',
'age', 'debt_to_income', 'dti_pct']
Stratification: on
```

Inspect outputs

[4]:

```
print("Train shape:", x_train.shape)
print("Test shape:", x_test.shape)
display(x_train.head())
```

```
Train shape: (3, 15)
Test shape: (2, 15)

   annual_income  employment_length  …  age_bin_55-64  age_bin_65+
2      -0.392232          -1.297771  …          False        False
0      -0.980581           0.162221  …          False        False
3       1.372813           1.135550  …          False        False

[3 rows x 15 columns]
```