Group 1 Final Project Work

Specific data on our dataset

Maternal Health Risk Dataset Summary

Shape: 808 records × 7 columns

Columns:

- `Age`
- `SystolicBP` (Systolic Blood Pressure)
- `DiastolicBP` (Diastolic Blood Pressure)
- `BS` (Blood Sugar level)
- `BodyTemp` (Body Temperature, °F)
- `HeartRate` (Heart Rate, bpm)
- `RiskLevel` (Target: maternal health risk category)

First 5 Records

Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate	RiskLevel
25	130	80	15.0	98.0	86	high risk
35	140	90	13.0	98.0	70	high risk
29	90	70	8.0	100.0	80	high risk
30	140	85	7.0	98.0	70	high risk
35	120	60	6.1	98.0	76	low risk

Summary Statistics

- **Age:** 10–70 years (mean = 30.6, std = 13.9)
- **SystolicBP:** 70–160 mmHg (mean = 113, std = 19.9)
- **DiastolicBP:** 49–100 mmHg (mean = 77.5, std = 14.8)
- **BS:** 6–19 mmol/L (mean = 9.26, std = 3.62)
- **BodyTemp:** 98–103 °F (mean = 98.6, std = 1.39)
- **HeartRate:** 7–90 bpm (mean = 74.3, std = 8.82)

Target Variable: RiskLevel

- **Low risk:** 478 records (~59.2%)

- **High risk:** 330 records (~40.8%)
- **Medium risk:** Not present in this dataset version

Note: The dataset is binary-labeled (low vs. high risk), so if a 3-class model (low/mid/high) is needed, additional data preprocessing or augmentation may be required.

Week 3 - Training and Feature Engineering

Environment (auto role + auto bucket) and constants

In [2]:
```
# This notebook:
#   • does EDA and writes plots/summary
#   • engineers features
#   • creates stratified splits: train(40%), val(10%), test(10%), production(40%)
#   • uploads artifacts to S3 (auto default bucket)
#   • creates OFFLINE Feature Store groups (auto execution role)
#   • writes a tracker update (JSON + Markdown)
#
# NO MANUAL SETTINGS: bucket/role are auto-detected from your Studio kernel.

import os, json, time
from pathlib import Path

import boto3
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler

import sagemaker
from sagemaker import get_execution_role
from sagemaker.session import Session

plt.rcParams["figure.dpi"] = 120

# Paths
LOCAL_DATA_PATH = Path("Maternal_Risk.csv")          # change only if your CSV is e
ARTIFACTS_DIR   = Path("week3_outputs")
ARTIFACTS_DIR.mkdir(parents=True, exist_ok=True)

# AWS context (auto)
boto_sess  = boto3.session.Session()
region     = boto_sess.region_name
sm_session = Session(boto_sess)
role       = get_execution_role()                    # auto from Studio kernel
bucket     = sm_session.default_bucket()             # auto default bucket

# Unique ids (prevent FG name collisions + make runs auditable)
RUN_ID    = time.strftime("%Y%m%d-%H%M%S")
S3_PREFIX = f"aai540/maternal-risk/week3/{RUN_ID}"

print("Region:", region)
```

```
print("Role:  ", role)
print("S3:    ", f"s3://{bucket}/{S3_PREFIX}")
```

```
sagemaker.config INFO - Not applying SDK defaults from location: /etc/xdg/sagemaker/
config.yaml
sagemaker.config INFO - Not applying SDK defaults from location: /home/sagemaker-use
r/.config/sagemaker/config.yaml
Region: us-east-1
Role:   arn:aws:iam::533267301342:role/LabRole
S3:     s3://sagemaker-us-east-1-533267301342/aai540/maternal-risk/week3/20250925-10
2647
```

Load data + lightweight EDA (plots + JSON summary)

In [3]:
```
assert LOCAL_DATA_PATH.exists(), f"Missing: {LOCAL_DATA_PATH.resolve()}"
df = pd.read_csv(LOCAL_DATA_PATH)

assert "RiskLevel" in df.columns, "Expected target column 'RiskLevel'."
print("Shape:", df.shape)
print("Columns:", list(df.columns))

# EDA summary (for tracker)
eda_summary = {
    "rows": int(df.shape[0]),
    "cols": int(df.shape[1]),
    "columns": df.columns.tolist(),
    "dtypes": {c: str(t) for c, t in df.dtypes.items()},
    "missing_counts": df.isna().sum().to_dict(),
    "class_counts": df["RiskLevel"].value_counts().to_dict(),
}
with open(ARTIFACTS_DIR / "eda_summary.json", "w") as f:
    json.dump(eda_summary, f, indent=2)

# Simple plots (defaults only; no custom colors)
df["RiskLevel"].value_counts().plot(kind="bar"); plt.title("Class Distribution")
plt.tight_layout(); plt.savefig(ARTIFACTS_DIR / "chart_class_distribution.png"); pl

df["Age"].plot(kind="hist", bins=20); plt.title("Age Distribution")
plt.tight_layout(); plt.savefig(ARTIFACTS_DIR / "chart_age_hist.png"); plt.clf()

plt.boxplot([df["SystolicBP"], df["DiastolicBP"]], labels=["SystolicBP","DiastolicB
plt.title("Blood Pressure Boxplots"); plt.tight_layout()
plt.savefig(ARTIFACTS_DIR / "chart_bp_box.png"); plt.clf()

num_cols = df.select_dtypes(include=[np.number]).columns
corr = df[num_cols].corr()
plt.imshow(corr, interpolation="nearest")
plt.xticks(range(len(num_cols)), num_cols, rotation=45, ha="right")
plt.yticks(range(len(num_cols)), num_cols); plt.colorbar(); plt.title("Correlation
plt.tight_layout(); plt.savefig(ARTIFACTS_DIR / "chart_corr_heatmap.png"); plt.clf(

print("EDA done →", ARTIFACTS_DIR)
```

```
Shape: (808, 7)
Columns: ['Age', 'SystolicBP', 'DiastolicBP', 'BS', 'BodyTemp', 'HeartRate', 'RiskLe
vel']
```

```
/tmp/ipykernel_215/2897146882.py:27: MatplotlibDeprecationWarning: The 'labels' para
meter of boxplot() has been renamed 'tick_labels' since Matplotlib 3.9; support for
the old name will be dropped in 3.11.
  plt.boxplot([df["SystolicBP"], df["DiastolicBP"]], labels=["SystolicBP","Diastolic
BP"])
EDA done → week3_outputs
<Figure size 768x576 with 0 Axes>
```

Feature engineering (clinically-motivated features + z-scaling)

In [4]:
```
# We derive simple vitals-based features and also add z-scaled versions for linear

X = df.copy()

# Clinically motivated derived features
X["PulsePressure"]     = X["SystolicBP"] - X["DiastolicBP"]
X["SBP_to_DBP"]        = X["SystolicBP"] / (X["DiastolicBP"].replace(0, np.nan))
X["Fever"]             = (X["BodyTemp"] > 99.5).astype(int)
X["Tachycardia"]       = (X["HeartRate"] >= 100).astype(int)
X["HypertensionFlag"]  = ((X["SystolicBP"] >= 140) | (X["DiastolicBP"] >= 90)).asty

# Optional standardization for linear models
cont = ["Age","SystolicBP","DiastolicBP","BS","BodyTemp","HeartRate","PulsePressure
X[[f"z_{c}" for c in cont]] = StandardScaler().fit_transform(X[cont])

# Label encoding (binary in this dataset)
label_map = {"low risk": 0, "high risk": 1}
if set(df["RiskLevel"].unique()) == set(label_map):
    y = df["RiskLevel"].map(label_map)
else:
    cats = sorted(df["RiskLevel"].unique())
    label_map = {v:i for i,v in enumerate(cats)}
    y = df["RiskLevel"].map(label_map)

with open(ARTIFACTS_DIR / "label_map.json", "w") as f:
    json.dump(label_map, f, indent=2)

X_no_target = X.drop(columns=["RiskLevel"])
engineered_full = pd.concat([X_no_target, y.rename("label")], axis=1)
engineered_full.to_csv(ARTIFACTS_DIR / "maternal_features_full.csv", index=False)

print("Feature engineering done.")
```

```
Feature engineering done.
```

Stratified splits: 40% prod, 40% train, 10% val, 10% test

In [5]:
```
# We first carve out 40% as "production" holdout for future batch inference/monitor
# The remaining 60% --> train (40%), val (10%), test (10%) of the original dataset.

from sklearn.model_selection import train_test_split

# 40% set aside for future batch inference/monitoring
X_tmp, X_prod, y_tmp, y_prod = train_test_split(
    X_no_target, y, test_size=0.40, random_state=42, stratify=y
```

```
)
# remaining 60% -> 40/10/10
X_train, X_rem, y_train, y_rem = train_test_split(
    X_tmp, y_tmp, test_size=(1/3), random_state=42, stratify=y_tmp
)
X_val, X_test, y_val, y_test = train_test_split(
    X_rem, y_rem, test_size=0.5, random_state=42, stratify=y_rem
)

def _save(name, Xd, yd):
    out = Xd.copy(); out["label"] = yd.values
    out.to_csv(ARTIFACTS_DIR / f"{name}.csv", index=False)
    return out

train_df = _save("train",      X_train, y_train)
val_df   = _save("val",        X_val,   y_val)
test_df  = _save("test",       X_test,  y_test)
prod_df  = _save("production", X_prod,  y_prod)

print({"train":len(train_df), "val":len(val_df), "test":len(test_df), "production":
```

```
{'train': 322, 'val': 81, 'test': 81, 'production': 324}
```

Upload artifacts to S3 (no manual bucket)

In [6]:

```
# Upload the CSVs, label map, EDA summary, and figures to your default bucket/prefi

s3 = boto3.client("s3")

def s3_upload(local: Path, key: str):
    s3.upload_file(str(local), bucket, f"{S3_PREFIX}/{key}")
    print("Uploaded", f"s3://{bucket}/{S3_PREFIX}/{key}")

# CSVs + summaries
for fname in ["train.csv","val.csv","test.csv","production.csv",
              "maternal_features_full.csv","label_map.json","eda_summary.json"]:
    s3_upload(ARTIFACTS_DIR / fname, fname)

# Plots
for fname in ["chart_class_distribution.png","chart_age_hist.png","chart_bp_box.png
    s3_upload(ARTIFACTS_DIR / fname, f"figures/{fname}")
```

```
Uploaded s3://sagemaker-us-east-1-533267301342/aai540/maternal-risk/week3/20250925-102647/train.csv
Uploaded s3://sagemaker-us-east-1-533267301342/aai540/maternal-risk/week3/20250925-102647/val.csv
Uploaded s3://sagemaker-us-east-1-533267301342/aai540/maternal-risk/week3/20250925-102647/test.csv
Uploaded s3://sagemaker-us-east-1-533267301342/aai540/maternal-risk/week3/20250925-102647/production.csv
Uploaded s3://sagemaker-us-east-1-533267301342/aai540/maternal-risk/week3/20250925-102647/maternal_features_full.csv
Uploaded s3://sagemaker-us-east-1-533267301342/aai540/maternal-risk/week3/20250925-102647/label_map.json
Uploaded s3://sagemaker-us-east-1-533267301342/aai540/maternal-risk/week3/20250925-102647/eda_summary.json
Uploaded s3://sagemaker-us-east-1-533267301342/aai540/maternal-risk/week3/20250925-102647/figures/chart_class_distribution.png
Uploaded s3://sagemaker-us-east-1-533267301342/aai540/maternal-risk/week3/20250925-102647/figures/chart_age_hist.png
Uploaded s3://sagemaker-us-east-1-533267301342/aai540/maternal-risk/week3/20250925-102647/figures/chart_bp_box.png
Uploaded s3://sagemaker-us-east-1-533267301342/aai540/maternal-risk/week3/20250925-102647/figures/chart_corr_heatmap.png
```

Sanitize column names (Feature Store regex) & write sanitized splits

In [7]:

```python
# FS rules: names must be letters/numbers/hyphens only; must start with alnum; <=64

def sanitize_col(name: str) -> str:
    if name == "SBP_to_DBP": name = "SBPtoDBP"   # preserve meaning
    if name.startswith("z_"): name = "z" + name[2:]
    name = name.replace("_", "")
    name = "".join(ch for ch in name if ch.isalnum() or ch == "-")
    if not name or not name[0].isalnum(): name = "f" + name
    return name[:64]

def sanitize_df_cols(df: pd.DataFrame) -> pd.DataFrame:
    newcols, seen = [], set()
    for c in df.columns:
        s = sanitize_col(c)
        if s in seen:
            i, base = 2, s
            while f"{base}{i}" in seen: i += 1
            s = f"{base}{i}"
        newcols.append(s); seen.add(s)
    out = df.copy(); out.columns = newcols
    return out

label_col = "label"
def sanitize_split(df):
    feats = df.drop(columns=[label_col])
    feats = sanitize_df_cols(feats)
    feats[label_col] = df[label_col].values
    return feats

train_s = sanitize_split(train_df); train_s.to_csv(ARTIFACTS_DIR/"train_sanitized.c
```

```
val_s   = sanitize_split(val_df);   val_s.to_csv(ARTIFACTS_DIR/"val_sanitized.csv",
test_s  = sanitize_split(test_df);  test_s.to_csv(ARTIFACTS_DIR/"test_sanitized.csv
prod_s  = sanitize_split(prod_df);  prod_s.to_csv(ARTIFACTS_DIR/"production_sanitiz

print("Sanitized splits saved.")
```

```
Sanitized splits saved.
```

Create & ingest Feature Store (OFFLINE, unique names per run)

In [8]:
```
import time
import boto3
from sagemaker.session import Session
from sagemaker.feature_store.feature_group import FeatureGroup
from sagemaker.feature_store.feature_definition import FeatureDefinition, FeatureTy

sm      = boto3.client("sagemaker")
session = Session(boto3.session.Session(region_name=region))

def ensure_id_time(df_in: pd.DataFrame) -> pd.DataFrame:
    df = df_in.copy()
    if "recordid" not in df.columns:
        df["recordid"] = range(1, len(df)+1)
    if "eventtime" not in df.columns:
        df["eventtime"] = pd.Timestamp.utcnow().isoformat()
    return df

def to_boto_feature_defs(df: pd.DataFrame):
    out = []
    for c, d in df.dtypes.items():
        if c == "eventtime":
            t = "String"
        elif pd.api.types.is_integer_dtype(d):
            t = "Integral"
        elif pd.api.types.is_float_dtype(d):
            t = "Fractional"
        else:
            t = "String"
        out.append({"FeatureName": c, "FeatureType": t})
    return out

def create_fg_boto3(name: str, df_local: pd.DataFrame, s3_uri: str):
    fdefs = to_boto_feature_defs(df_local)
    try:
        resp = sm.create_feature_group(
            FeatureGroupName=name,
            RecordIdentifierFeatureName="recordid",
            EventTimeFeatureName="eventtime",
            FeatureDefinitions=fdefs,
            OfflineStoreConfig={"S3StorageConfig": {"S3Uri": s3_uri}},
            OnlineStoreConfig={"EnableOnlineStore": False},
            RoleArn=role,
            Description=f"Maternal Health Risk – {name}",
        )
        return resp
    except sm.exceptions.ResourceInUse:
```

```
        # Already exists --> safe to reuse after we confirm it's Created
        return {"FeatureGroupArn": f"arn:aws:sagemaker:{region}:{boto3.client('sts'

def wait_fg_created(name: str, timeout_s: int = 900, poll_s: int = 10):
    start = time.time()
    last = ""
    while True:
        desc = sm.describe_feature_group(FeatureGroupName=name)
        status = desc.get("FeatureGroupStatus", "")
        if status == "Created":
            print(f"[READY] {name}")
            return desc
        if status == "CreateFailed":
            raise RuntimeError(f"{name} failed: {desc.get('FailureReason')}")
        if time.time() - start > timeout_s:
            raise TimeoutError(f"Timeout waiting for {name} (last status={status})"
        if status != last:
            print(f"Status {name}: {status}")
            last = status
        time.sleep(poll_s)

def create_and_ingest(name_base: str, df_local: pd.DataFrame):
    # unique FG names per run to avoid collisions
    name = f"{name_base}-{RUN_ID}"             # e.g., mhr-train-fg-20250920-154301
    assert "_" not in name, "FG name must not contain underscores."
    df_local = ensure_id_time(df_local)
    s3_uri   = f"s3://{bucket}/{S3_PREFIX}/feature-store/{name}"

    create_fg_boto3(name, df_local, s3_uri)
    wait_fg_created(name)

    fg = FeatureGroup(name=name, sagemaker_session=session)
    fg.load_feature_definitions(data_frame=df_local)    # make sure SDK knows schem
    fg.ingest(data_frame=df_local, max_workers=4, wait=True)
    print(f"[OK] Ingested {len(df_local)} rows → {name}")
    return name

# Load sanitized splits
train_s = pd.read_csv(ARTIFACTS_DIR/"train_sanitized.csv")
val_s   = pd.read_csv(ARTIFACTS_DIR/"val_sanitized.csv")
prod_s  = pd.read_csv(ARTIFACTS_DIR/"production_sanitized.csv")

# Create OFFLINE FGs with unique names
FG_TRAIN = create_and_ingest("mhr-train-fg", train_s)
FG_VAL   = create_and_ingest("mhr-val-fg",   val_s)
FG_BATCH = create_and_ingest("mhr-batch-fg", prod_s)

print("Feature Store complete:", FG_TRAIN, FG_VAL, FG_BATCH)
```

```
Status mhr-train-fg-20250925-102647: Creating
[READY] mhr-train-fg-20250925-102647
[OK] Ingested 322 rows → mhr-train-fg-20250925-102647
Status mhr-val-fg-20250925-102647: Creating
[READY] mhr-val-fg-20250925-102647
[OK] Ingested 81 rows → mhr-val-fg-20250925-102647
Status mhr-batch-fg-20250925-102647: Creating
[READY] mhr-batch-fg-20250925-102647
[OK] Ingested 324 rows → mhr-batch-fg-20250925-102647
Feature Store complete: mhr-train-fg-20250925-102647 mhr-val-fg-20250925-102647 mhr-batch-fg-20250925-102647
```

Tracker update (JSON + Markdown) and upload

In [9]:

```
tracker = {
    "run_id": RUN_ID,
    "s3_prefix": f"s3://{bucket}/{S3_PREFIX}",
    "dataset": {
        "rows": eda_summary["rows"], "cols": eda_summary["cols"],
        "class_counts": eda_summary["class_counts"],
        "dtypes": eda_summary["dtypes"],
        "missing": eda_summary["missing_counts"],
    },
    "splits": {
        "train_rows": len(train_df), "val_rows": len(val_df),
        "test_rows": len(test_df), "prod_rows": len(prod_df),
    },
    "feature_store_groups": [FG_TRAIN, FG_VAL, FG_BATCH],
}
with open(ARTIFACTS_DIR / "team_tracker_update_week3.json", "w") as f:
    json.dump(tracker, f, indent=2)

md = f"""# Week 3 Tracker — Maternal Health Risk (RUN: {RUN_ID})

**S3 prefix:** s3://{bucket}/{S3_PREFIX}

## Dataset
- Rows: {eda_summary['rows']} | Cols: {eda_summary['cols']}
- Classes: {eda_summary['class_counts']}

## Splits
- Train: {len(train_df)} (~40%)
- Val:   {len(val_df)} (~10%)
- Test:  {len(test_df)} (~10%)
- Prod:  {len(prod_df)} (~40%)

## Feature Store (offline)
- {FG_TRAIN}
- {FG_VAL}
- {FG_BATCH}
"""
with open(ARTIFACTS_DIR / "team_tracker_update_week3.md", "w") as f:
    f.write(md)

# Upload tracker docs
```

```
s3 = boto3.client("s3")
s3.upload_file(str(ARTIFACTS_DIR/"team_tracker_update_week3.json"), bucket, f"{S3_P
s3.upload_file(str(ARTIFACTS_DIR/"team_tracker_update_week3.md"),   bucket, f"{S3_P

print("Tracker written & uploaded.")
```

```
Tracker written & uploaded.
```

In [10]:
```
# View

import boto3, json
s3 = boto3.client("s3")
obj = s3.get_object(Bucket=bucket, Key=f"{S3_PREFIX}/team_tracker_update_week3.json
tracker = json.load(obj["Body"])
tracker
```

Out[10]:

```
{'run_id': '20250925-102647',
 's3_prefix': 's3://sagemaker-us-east-1-533267301342/aai540/maternal-risk/week3/20
250925-102647',
 'dataset': {'rows': 808,
  'cols': 7,
  'class_counts': {'low risk': 478, 'high risk': 330},
  'dtypes': {'Age': 'int64',
   'SystolicBP': 'int64',
   'DiastolicBP': 'int64',
   'BS': 'float64',
   'BodyTemp': 'float64',
   'HeartRate': 'int64',
   'RiskLevel': 'object'},
  'missing': {'Age': 0,
   'SystolicBP': 0,
   'DiastolicBP': 0,
   'BS': 0,
   'BodyTemp': 0,
   'HeartRate': 0,
   'RiskLevel': 0}},
 'splits': {'train_rows': 322,
  'val_rows': 81,
  'test_rows': 81,
  'prod_rows': 324},
 'feature_store_groups': ['mhr-train-fg-20250925-102647',
  'mhr-val-fg-20250925-102647',
  'mhr-batch-fg-20250925-102647']}
```


Week 4, Model Development and Deployment

WEEK 4: CONTEXT (auto settings; continues from Week 3)

In [11]:
```
import os, io, json, time, tarfile
from pathlib import Path
import boto3, sagemaker
from sagemaker import get_execution_role
from sagemaker.session import Session
import pandas as pd
```

```
import numpy as np

# Reuse Week-3 objects if they exist; otherwise, auto-init (no manual config)
try:
    bucket
    sm_session
    role
except NameError:
    boto_sess  = boto3.session.Session()
    sm_session = Session(boto_sess)
    role       = get_execution_role()
    bucket     = sm_session.default_bucket()

# Use the Week-3 S3 prefix if it's still in memory; otherwise pick the latest run
s3 = boto3.client("s3")
try:
    WEEK3_PREFIX = S3_PREFIX  # from Week 3 cells
except NameError:
    base = "aai540/maternal-risk/week3/"
    resp = s3.list_objects_v2(Bucket=bucket, Prefix=base, Delimiter="/")
    runs = [cp["Prefix"].rstrip("/") for cp in resp.get("CommonPrefixes", [])]
    assert runs, f"No Week-3 artifacts found under s3://{bucket}/{base}"
    WEEK3_PREFIX = sorted(runs)[-1]

# Create a unique Week-4 prefix
RUN_ID = time.strftime("%Y%m%d-%H%M%S")
W4_PREFIX = f"aai540/maternal-risk/week4/{RUN_ID}"

print("Using Week-3:", f"s3://{bucket}/{WEEK3_PREFIX}")
print("Writing Week-4:", f"s3://{bucket}/{W4_PREFIX}")
```

```
Using Week-3: s3://sagemaker-us-east-1-533267301342/aai540/maternal-risk/week3/20250
925-102647
Writing Week-4: s3://sagemaker-us-east-1-533267301342/aai540/maternal-risk/week4/202
50925-102857
```

Load Week-3 splits (train/val/test) from S3

In [12]:

```
# LOAD SPLITS

def read_csv_from_s3(key: str) -> pd.DataFrame:
    obj = s3.get_object(Bucket=bucket, Key=key)
    return pd.read_csv(io.BytesIO(obj["Body"].read()))

train = read_csv_from_s3(f"{WEEK3_PREFIX}/train.csv")
val   = read_csv_from_s3(f"{WEEK3_PREFIX}/val.csv")
test  = read_csv_from_s3(f"{WEEK3_PREFIX}/test.csv")

label_col = "label"
X_train, y_train = train.drop(columns=[label_col]), train[label_col]
X_val,   y_val   = val.drop(columns=[label_col]),   val[label_col]
X_test,  y_test  = test.drop(columns=[label_col]),  test[label_col]

print("Loaded:", train.shape, val.shape, test.shape)
print("Train label balance:", y_train.value_counts().to_dict())
```

```
Loaded: (322, 20) (81, 20) (81, 20)
Train label balance: {0: 190, 1: 132}
```

Benchmark model in SageMaker (very simple: Logistic Regression on 2 features)

In [13]:
```
from sagemaker.sklearn import SKLearn
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, precision_recall_fscore_support, roc_au
import numpy as np
from pathlib import Path
import pandas as pd

bm_feats = ["Age", "SystolicBP"]
assert all(f in X_train.columns for f in bm_feats), "Missing expected features for

# Save small CSVs for the training job (filenames do NOT matter in SageMaker)
w4_local = Path("w4_benchmark"); w4_local.mkdir(exist_ok=True)
pd.concat([X_train[bm_feats], y_train], axis=1).to_csv(w4_local/"train_benchmark.cs
pd.concat([X_val[bm_feats],   y_val],   axis=1).to_csv(w4_local/"val_benchmark.csv"

bm_train_s3 = sm_session.upload_data(str(w4_local/"train_benchmark.csv"), key_prefi
bm_val_s3   = sm_session.upload_data(str(w4_local/"val_benchmark.csv"),   key_prefi

# --- WEEK 4 change: entry script reads the FIRST *.csv in each channel dir via env
with open("baseline_train.py","w") as f:
    f.write("""
import os, glob, json, pathlib
import pandas as pd
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, precision_recall_fscore_support, roc_au

def first_csv_in(dir_path):
    files = sorted(glob.glob(os.path.join(dir_path, '*.csv')))
    assert files, f'No CSV found in {dir_path}'
    return files[0]

if __name__ == '__main__':
    # Channels provided by SageMaker
    train_dir = os.environ.get('SM_CHANNEL_TRAIN', '/opt/ml/input/data/train')
    val_dir   = os.environ.get('SM_CHANNEL_VAL',   '/opt/ml/input/data/val')
    model_dir = os.environ.get('SM_MODEL_DIR',     '/opt/ml/model')

    train_path = first_csv_in(train_dir)
    val_path   = first_csv_in(val_dir)

    df_tr = pd.read_csv(train_path)
    df_va = pd.read_csv(val_path)

    Xtr, ytr = df_tr[['Age','SystolicBP']], df_tr['label']
    Xva, yva = df_va[['Age','SystolicBP']], df_va['label']

    clf = LogisticRegression(max_iter=1000)
    clf.fit(Xtr, ytr)
```

```
    # Metrics on val
    pred  = clf.predict(Xva)
    proba = clf.predict_proba(Xva)[:,1]
    acc = accuracy_score(yva, pred)
    p, r, f1, _ = precision_recall_fscore_support(yva, pred, average='binary', zero
    try:
        auc = roc_auc_score(yva, proba)
    except ValueError:
        auc = float('nan')

    pathlib.Path(model_dir).mkdir(parents=True, exist_ok=True)
    import joblib
    joblib.dump(clf, os.path.join(model_dir, 'model.joblib'))
    with open(os.path.join(model_dir, 'metrics.json'), 'w') as f:
        json.dump({'accuracy':acc,'precision':p,'recall':r,'f1':f1,'roc_auc':auc},
""")

bm_est = SKLearn(
    entry_point="baseline_train.py",
    framework_version="1.2-1",     # keep your version; change only if your Studio
    role=role,
    instance_type="ml.m5.large",
    instance_count=1,
    sagemaker_session=sm_session,
)

bm_est.fit({"train": bm_train_s3, "val": bm_val_s3})
bm_model_artifact = bm_est.model_data
print("Baseline model artifact:", bm_model_artifact)

# Refit the same baseline locally for clean side-by-side metrics (unchanged)
bm_clf = LogisticRegression(max_iter=1000).fit(train[bm_feats], y_train)
bm_proba = bm_clf.predict_proba(test[bm_feats])[:,1]
bm_pred  = (bm_proba >= 0.5).astype(int)
bm_acc   = accuracy_score(y_test, bm_pred)
bm_p, bm_r, bm_f1, _ = precision_recall_fscore_support(y_test, bm_pred, average='bi
try:
    bm_auc = roc_auc_score(y_test, bm_proba)
except ValueError:
    bm_auc = float('nan')

baseline_metrics = {"accuracy":bm_acc,"precision":bm_p,"recall":bm_r,"f1":bm_f1,"ro
baseline_metrics
```

```
INFO:sagemaker:Creating training-job with name: sagemaker-scikit-learn-2025-09-25-10
-29-02-273
```

```
2025-09-25 10:29:03 Starting - Starting the training job...
2025-09-25 10:29:18 Starting - Preparing the instances for training...
2025-09-25 10:30:04 Downloading - Downloading the training image........./miniconda3/
lib/python3.9/site-packages/sagemaker_containers/_server.py:22: UserWarning: pkg_res
ources is deprecated as an API. See https://setuptools.pypa.io/en/latest/pkg_resourc
es.html. The pkg_resources package is slated for removal as early as 2025-11-30. Ref
rain from using this package or pin to Setuptools<81.
  import pkg_resources
2025-09-25 10:31:13,657 sagemaker-containers INFO     Imported framework sagemaker_s
klearn_container.training
2025-09-25 10:31:13,662 sagemaker-training-toolkit INFO     No GPUs detected (normal
if no gpus installed)
2025-09-25 10:31:13,664 sagemaker-training-toolkit INFO     No Neurons detected (nor
mal if no neurons installed)
2025-09-25 10:31:13,682 sagemaker_sklearn_container.training INFO     Invoking user
training script.
2025-09-25 10:31:13,968 sagemaker-training-toolkit INFO     No GPUs detected (normal
if no gpus installed)
2025-09-25 10:31:13,971 sagemaker-training-toolkit INFO     No Neurons detected (nor
mal if no neurons installed)
2025-09-25 10:31:13,990 sagemaker-training-toolkit INFO     No GPUs detected (normal
if no gpus installed)
2025-09-25 10:31:13,992 sagemaker-training-toolkit INFO     No Neurons detected (nor
mal if no neurons installed)
2025-09-25 10:31:14,010 sagemaker-training-toolkit INFO     No GPUs detected (normal
if no gpus installed)
2025-09-25 10:31:14,013 sagemaker-training-toolkit INFO     No Neurons detected (nor
mal if no neurons installed)
2025-09-25 10:31:14,028 sagemaker-training-toolkit INFO     Invoking user script
Training Env:
{
    "additional_framework_parameters": {},
    "channel_input_dirs": {
        "train": "/opt/ml/input/data/train",
        "val": "/opt/ml/input/data/val"
    },
    "current_host": "algo-1",
    "current_instance_group": "homogeneousCluster",
    "current_instance_group_hosts": [
        "algo-1"
    ],
    "current_instance_type": "ml.m5.large",
    "distribution_hosts": [],
    "distribution_instance_groups": [],
    "framework_module": "sagemaker_sklearn_container.training:main",
    "hosts": [
        "algo-1"
    ],
    "hyperparameters": {},
    "input_config_dir": "/opt/ml/input/config",
    "input_data_config": {
        "train": {
            "TrainingInputMode": "File",
            "S3DistributionType": "FullyReplicated",
            "RecordWrapperType": "None"
        },
```

```
            "val": {
                "TrainingInputMode": "File",
                "S3DistributionType": "FullyReplicated",
                "RecordWrapperType": "None"
            }
        },
        "input_dir": "/opt/ml/input",
        "instance_groups": [
            "homogeneousCluster"
        ],
        "instance_groups_dict": {
            "homogeneousCluster": {
                "instance_group_name": "homogeneousCluster",
                "instance_type": "ml.m5.large",
                "hosts": [
                    "algo-1"
                ]
            }
        },
        "is_hetero": false,
        "is_master": true,
        "is_modelparallel_enabled": null,
        "is_smddpmprun_installed": false,
        "is_smddprun_installed": false,
        "job_name": "sagemaker-scikit-learn-2025-09-25-10-29-02-273",
        "log_level": 20,
        "master_hostname": "algo-1",
        "model_dir": "/opt/ml/model",
        "module_dir": "s3://sagemaker-us-east-1-533267301342/sagemaker-scikit-learn-2025
    -09-25-10-29-02-273/source/sourcedir.tar.gz",
        "module_name": "baseline_train",
        "network_interface_name": "eth0",
        "num_cpus": 2,
        "num_gpus": 0,
        "num_neurons": 0,
        "output_data_dir": "/opt/ml/output/data",
        "output_dir": "/opt/ml/output",
        "output_intermediate_dir": "/opt/ml/output/intermediate",
        "resource_config": {
            "current_host": "algo-1",
            "current_instance_type": "ml.m5.large",
            "current_group_name": "homogeneousCluster",
            "hosts": [
                "algo-1"
            ],
            "instance_groups": [
                {
                    "instance_group_name": "homogeneousCluster",
                    "instance_type": "ml.m5.large",
                    "hosts": [
                        "algo-1"
                    ]
                }
            ],
            "network_interface_name": "eth0",
            "topology": null
```

```
    },
    "user_entry_point": "baseline_train.py"
}
Environment variables:
SM_HOSTS=["algo-1"]
SM_NETWORK_INTERFACE_NAME=eth0
SM_HPS={}
SM_USER_ENTRY_POINT=baseline_train.py
SM_FRAMEWORK_PARAMS={}
SM_RESOURCE_CONFIG={"current_group_name":"homogeneousCluster","current_host":"algo-
1","current_instance_type":"ml.m5.large","hosts":["algo-1"],"instance_groups":[{"hos
ts":["algo-1"],"instance_group_name":"homogeneousCluster","instance_type":"ml.m5.lar
ge"}],"network_interface_name":"eth0","topology":null}
SM_INPUT_DATA_CONFIG={"train":{"RecordWrapperType":"None","S3DistributionType":"Full
yReplicated","TrainingInputMode":"File"},"val":{"RecordWrapperType":"None","S3Distri
butionType":"FullyReplicated","TrainingInputMode":"File"}}
SM_OUTPUT_DATA_DIR=/opt/ml/output/data
SM_CHANNELS=["train","val"]
SM_CURRENT_HOST=algo-1
SM_CURRENT_INSTANCE_TYPE=ml.m5.large
SM_CURRENT_INSTANCE_GROUP=homogeneousCluster
SM_CURRENT_INSTANCE_GROUP_HOSTS=["algo-1"]
SM_INSTANCE_GROUPS=["homogeneousCluster"]
SM_INSTANCE_GROUPS_DICT={"homogeneousCluster":{"hosts":["algo-1"],"instance_group_na
me":"homogeneousCluster","instance_type":"ml.m5.large"}}
SM_DISTRIBUTION_INSTANCE_GROUPS=[]
SM_IS_HETERO=false
SM_MODULE_NAME=baseline_train
SM_LOG_LEVEL=20
SM_FRAMEWORK_MODULE=sagemaker_sklearn_container.training:main
SM_INPUT_DIR=/opt/ml/input
SM_INPUT_CONFIG_DIR=/opt/ml/input/config
SM_OUTPUT_DIR=/opt/ml/output
SM_NUM_CPUS=2
SM_NUM_GPUS=0
SM_NUM_NEURONS=0
SM_MODEL_DIR=/opt/ml/model
SM_MODULE_DIR=s3://sagemaker-us-east-1-533267301342/sagemaker-scikit-learn-2025-09-2
5-10-29-02-273/source/sourcedir.tar.gz
SM_TRAINING_ENV={"additional_framework_parameters":{},"channel_input_dirs":{"trai
n":"/opt/ml/input/data/train","val":"/opt/ml/input/data/val"},"current_host":"algo-
1","current_instance_group":"homogeneousCluster","current_instance_group_hosts":["al
go-1"],"current_instance_type":"ml.m5.large","distribution_hosts":[],"distribution_i
nstance_groups":[],"framework_module":"sagemaker_sklearn_container.training:main","h
osts":["algo-1"],"hyperparameters":{},"input_config_dir":"/opt/ml/input/config","inp
ut_data_config":{"train":{"RecordWrapperType":"None","S3DistributionType":"FullyRepl
icated","TrainingInputMode":"File"},"val":{"RecordWrapperType":"None","S3Distributio
nType":"FullyReplicated","TrainingInputMode":"File"}},"input_dir":"/opt/ml/input","i
nstance_groups":["homogeneousCluster"],"instance_groups_dict":{"homogeneousCluster":
{"hosts":["algo-1"],"instance_group_name":"homogeneousCluster","instance_type":"ml.m
5.large"}},"is_hetero":false,"is_master":true,"is_modelparallel_enabled":null,"is_sm
ddpmprun_installed":false,"is_smddprun_installed":false,"job_name":"sagemaker-scikit
-learn-2025-09-25-10-29-02-273","log_level":20,"master_hostname":"algo-1","model_di
r":"/opt/ml/model","module_dir":"s3://sagemaker-us-east-1-533267301342/sagemaker-sci
kit-learn-2025-09-25-10-29-02-273/source/sourcedir.tar.gz","module_name":"baseline_t
rain","network_interface_name":"eth0","num_cpus":2,"num_gpus":0,"num_neurons":0,"out
```

```
put_data_dir":"/opt/ml/output/data","output_dir":"/opt/ml/output","output_intermedia
te_dir":"/opt/ml/output/intermediate","resource_config":{"current_group_name":"homog
eneousCluster","current_host":"algo-1","current_instance_type":"ml.m5.large","host
s":["algo-1"],"instance_groups":[{"hosts":["algo-1"],"instance_group_name":"homogene
ousCluster","instance_type":"ml.m5.large"}],"network_interface_name":"eth0","topolog
y":null},"user_entry_point":"baseline_train.py"}
SM_USER_ARGS=[]
SM_OUTPUT_INTERMEDIATE_DIR=/opt/ml/output/intermediate
SM_CHANNEL_TRAIN=/opt/ml/input/data/train
SM_CHANNEL_VAL=/opt/ml/input/data/val
PYTHONPATH=/opt/ml/code:/miniconda3/bin:/miniconda3/lib/python39.zip:/miniconda3/li
b/python3.9:/miniconda3/lib/python3.9/lib-dynload:/miniconda3/lib/python3.9/site-pac
kages:/miniconda3/lib/python3.9/site-packages/setuptools/_vendor
Invoking script with the following command:
/miniconda3/bin/python baseline_train.py
2025-09-25 10:31:14,029 sagemaker-training-toolkit INFO     Exceptions not imported
for SageMaker Debugger as it is not installed.
2025-09-25 10:31:14,030 sagemaker-training-toolkit INFO     Exceptions not imported
for SageMaker TF as Tensorflow is not installed.
2025-09-25 10:31:14,948 sagemaker-containers INFO     Reporting training SUCCESS

2025-09-25 10:31:34 Training - Training image download completed. Training in progre
ss.
2025-09-25 10:31:34 Uploading - Uploading generated training model
2025-09-25 10:31:34 Completed - Training job completed
Training seconds: 114
Billable seconds: 114
Baseline model artifact: s3://sagemaker-us-east-1-533267301342/sagemaker-scikit-lear
n-2025-09-25-10-29-02-273/output/model.tar.gz
```

Out[13]:
```
{'accuracy': 0.7654320987654321,
 'precision': 0.71875,
 'recall': 0.696969696969697,
 'f1': 0.7076923076923077,
 'roc_auc': 0.790719696969697}
```

Main model in SageMaker (XGBoost built-in)

In [14]:
```
from pathlib import Path
import sagemaker
from sagemaker.estimator import Estimator
from sagemaker.inputs import TrainingInput

# Reorder so label is first (as XGBoost CSV expects)
def reorder_for_xgb(df):
    cols = [label_col] + [c for c in df.columns if c != label_col]
    return df[cols]

xgb_train_local = reorder_for_xgb(train)
xgb_val_local   = reorder_for_xgb(val)

xgb_train_path  = Path("w4_xgb_train.csv")
xgb_val_path    = Path("w4_xgb_val.csv")

# IMPORTANT: no header for built-in XGB CSV
xgb_train_local.to_csv(xgb_train_path, index=False, header=False)
```

```
xgb_val_local.to_csv(xgb_val_path,   index=False, header=False)

s3_xgb_train = sm_session.upload_data(str(xgb_train_path), key_prefix=f"{W4_PREFIX}
s3_xgb_val   = sm_session.upload_data(str(xgb_val_path),   key_prefix=f"{W4_PREFIX}

# Pick an available built-in XGBoost image
def get_xgb_image():
    for ver in ["1.7-1", "1.5-1", "1.3-1"]:
        try:
            return sagemaker.image_uris.retrieve("xgboost", sm_session.boto_region_
        except Exception as e:
            print(f"xgboost {ver} not available, trying next… ({e})")
    raise RuntimeError("No compatible built-in XGBoost image found.")

xgb_image_uri = get_xgb_image()

xgb_est = Estimator(
    image_uri=xgb_image_uri,
    role=role,
    instance_count=1,
    instance_type="ml.m5.large",
    sagemaker_session=sm_session,
    hyperparameters={
        # CSV mode: label must be first column (we already set that)
        "objective": "binary:logistic",
        "eval_metric": "auc",
        "max_depth": 5,
        "eta": 0.2,
        "min_child_weight": 1,
        "subsample": 0.8,
        "colsample_bytree": 0.8,
        "num_round": 200,
        "verbosity": 1,
    },
)

# WEEK 4 change: declare CSV content type on both channels
train_input = TrainingInput(s3_data=s3_xgb_train, content_type="text/csv")
val_input   = TrainingInput(s3_data=s3_xgb_val,   content_type="text/csv")

xgb_est.fit({"train": train_input, "validation": val_input}, wait=True)

xgb_model_artifact = xgb_est.model_data
print("XGBoost model artifact:", xgb_model_artifact)
```

```
INFO:sagemaker.image_uris:Ignoring unnecessary instance type: None.
INFO:sagemaker.telemetry.telemetry_logging:SageMaker Python SDK will collect telemet
ry to help us better understand our user's needs, diagnose issues, and deliver addit
ional features.
To opt out of telemetry, please disable via TelemetryOptOut parameter in SDK default
s config. For more information, refer to https://sagemaker.readthedocs.io/en/stable/
overview.html#configuring-and-using-defaults-with-the-sagemaker-python-sdk.
INFO:sagemaker:Creating training-job with name: sagemaker-xgboost-2025-09-25-10-31-5
6-289
```

```
2025-09-25 10:32:01 Starting - Starting the training job...
2025-09-25 10:32:16 Starting - Preparing the instances for training...
2025-09-25 10:32:36 Downloading - Downloading input data...
2025-09-25 10:33:16 Downloading - Downloading the training image......
2025-09-25 10:34:27 Training - Training image download completed. Training in progre
ss../miniconda3/lib/python3.9/site-packages/sagemaker_containers/_server.py:22: User
Warning: pkg_resources is deprecated as an API. See https://setuptools.pypa.io/en/la
test/pkg_resources.html. The pkg_resources package is slated for removal as early as
2025-11-30. Refrain from using this package or pin to Setuptools<81.
  import pkg_resources
[2025-09-25 10:34:31.168 ip-10-0-120-196.ec2.internal:7 INFO utils.py:28] RULE_JOB_S
TOP_SIGNAL_FILENAME: None
[2025-09-25 10:34:31.238 ip-10-0-120-196.ec2.internal:7 INFO profiler_config_parser.
py:111] User has disabled profiler.
[2025-09-25:10:34:31:INFO] Imported framework sagemaker_xgboost_container.training
[2025-09-25:10:34:31:INFO] Failed to parse hyperparameter eval_metric value auc to J
son.
Returning the value itself
[2025-09-25:10:34:31:INFO] Failed to parse hyperparameter objective value binary:log
istic to Json.
Returning the value itself
[2025-09-25:10:34:31:INFO] No GPUs detected (normal if no gpus installed)
[2025-09-25:10:34:31:INFO] Running XGBoost Sagemaker in algorithm mode
[2025-09-25:10:34:31:INFO] Determined 0 GPU(s) available on the instance.
[2025-09-25:10:34:31:INFO] Determined delimiter of CSV input is ','
[2025-09-25:10:34:31:INFO] Determined delimiter of CSV input is ','
[2025-09-25:10:34:31:INFO] File path /opt/ml/input/data/train of input files
[2025-09-25:10:34:31:INFO] Making smlinks from folder /opt/ml/input/data/train to fo
lder /tmp/sagemaker_xgboost_input_data
[2025-09-25:10:34:31:INFO] creating symlink between Path /opt/ml/input/data/train/w4
_xgb_train.csv and destination /tmp/sagemaker_xgboost_input_data/w4_xgb_train.csv-76
22590553926284355
[2025-09-25:10:34:31:INFO] files path: /tmp/sagemaker_xgboost_input_data
[2025-09-25:10:34:31:INFO] Determined delimiter of CSV input is ','
[2025-09-25:10:34:31:INFO] File path /opt/ml/input/data/validation of input files
[2025-09-25:10:34:31:INFO] Making smlinks from folder /opt/ml/input/data/validation
to folder /tmp/sagemaker_xgboost_input_data
[2025-09-25:10:34:31:INFO] creating symlink between Path /opt/ml/input/data/validati
on/w4_xgb_val.csv and destination /tmp/sagemaker_xgboost_input_data/w4_xgb_val.csv81
85667522940232075
[2025-09-25:10:34:31:INFO] files path: /tmp/sagemaker_xgboost_input_data
[2025-09-25:10:34:31:INFO] Determined delimiter of CSV input is ','
[2025-09-25:10:34:31:INFO] Single node training.
[2025-09-25:10:34:31:INFO] Train matrix has 322 rows and 19 columns
[2025-09-25:10:34:31:INFO] Validation matrix has 81 rows
[2025-09-25 10:34:31.606 ip-10-0-120-196.ec2.internal:7 INFO json_config.py:92] Crea
ting hook from json_config at /opt/ml/input/config/debughookconfig.json.
[2025-09-25 10:34:31.607 ip-10-0-120-196.ec2.internal:7 INFO hook.py:206] tensorboar
d_dir has not been set for the hook. SMDebug will not be exporting tensorboard summa
ries.
[2025-09-25 10:34:31.607 ip-10-0-120-196.ec2.internal:7 INFO hook.py:259] Saving to
/opt/ml/output/tensors
[2025-09-25 10:34:31.607 ip-10-0-120-196.ec2.internal:7 INFO state_store.py:77] The
checkpoint config file /opt/ml/input/config/checkpointconfig.json does not exist.
[2025-09-25:10:34:31:INFO] Debug hook created from config
[2025-09-25 10:34:31.610 ip-10-0-120-196.ec2.internal:7 INFO hook.py:427] Monitoring
```

```
the collections: metrics
[2025-09-25 10:34:31.614 ip-10-0-120-196.ec2.internal:7 INFO hook.py:491] Hook is wr
iting from the hook with pid: 7
[0]#011train-auc:0.96288#011validation-auc:0.92803
[1]#011train-auc:0.96423#011validation-auc:0.92803
[2]#011train-auc:0.97907#011validation-auc:0.94760
[3]#011train-auc:0.98569#011validation-auc:0.95960
[4]#011train-auc:0.98728#011validation-auc:0.95960
[5]#011train-auc:0.99244#011validation-auc:0.98359
[6]#011train-auc:0.99169#011validation-auc:0.98232
[7]#011train-auc:0.99256#011validation-auc:0.98359
[8]#011train-auc:0.99340#011validation-auc:0.98359
[9]#011train-auc:0.99402#011validation-auc:0.98201
[10]#011train-auc:0.99400#011validation-auc:0.97917
[11]#011train-auc:0.99468#011validation-auc:0.98106
[12]#011train-auc:0.99482#011validation-auc:0.98106
[13]#011train-auc:0.99490#011validation-auc:0.98106
[14]#011train-auc:0.99502#011validation-auc:0.98232
[15]#011train-auc:0.99535#011validation-auc:0.98485
[16]#011train-auc:0.99559#011validation-auc:0.97854
[17]#011train-auc:0.99623#011validation-auc:0.97854
[18]#011train-auc:0.99643#011validation-auc:0.97727
[19]#011train-auc:0.99699#011validation-auc:0.97790
[20]#011train-auc:0.99711#011validation-auc:0.98169
[21]#011train-auc:0.99735#011validation-auc:0.98422
[22]#011train-auc:0.99727#011validation-auc:0.98422
[23]#011train-auc:0.99719#011validation-auc:0.98485
[24]#011train-auc:0.99731#011validation-auc:0.98359
[25]#011train-auc:0.99767#011validation-auc:0.98232
[26]#011train-auc:0.99767#011validation-auc:0.98295
[27]#011train-auc:0.99803#011validation-auc:0.98169
[28]#011train-auc:0.99815#011validation-auc:0.98232
[29]#011train-auc:0.99815#011validation-auc:0.98232
[30]#011train-auc:0.99831#011validation-auc:0.98295
[31]#011train-auc:0.99831#011validation-auc:0.98359
[32]#011train-auc:0.99850#011validation-auc:0.98422
[33]#011train-auc:0.99831#011validation-auc:0.98359
[34]#011train-auc:0.99839#011validation-auc:0.98422
[35]#011train-auc:0.99831#011validation-auc:0.98485
[36]#011train-auc:0.99854#011validation-auc:0.98864
[37]#011train-auc:0.99854#011validation-auc:0.98611
[38]#011train-auc:0.99854#011validation-auc:0.98801
[39]#011train-auc:0.99854#011validation-auc:0.98674
[40]#011train-auc:0.99878#011validation-auc:0.98737
[41]#011train-auc:0.99878#011validation-auc:0.98737
[42]#011train-auc:0.99878#011validation-auc:0.98864
[43]#011train-auc:0.99878#011validation-auc:0.98674
[44]#011train-auc:0.99878#011validation-auc:0.98611
[45]#011train-auc:0.99878#011validation-auc:0.98611
[46]#011train-auc:0.99886#011validation-auc:0.98548
[47]#011train-auc:0.99874#011validation-auc:0.98611
[48]#011train-auc:0.99886#011validation-auc:0.98737
[49]#011train-auc:0.99886#011validation-auc:0.98737
[50]#011train-auc:0.99886#011validation-auc:0.98737
[51]#011train-auc:0.99890#011validation-auc:0.98801
[52]#011train-auc:0.99890#011validation-auc:0.98801
```

```
[53]#011train-auc:0.99890#011validation-auc:0.98801
[54]#011train-auc:0.99890#011validation-auc:0.98737
[55]#011train-auc:0.99878#011validation-auc:0.98737
[56]#011train-auc:0.99878#011validation-auc:0.98737
[57]#011train-auc:0.99878#011validation-auc:0.98737
[58]#011train-auc:0.99890#011validation-auc:0.98737
[59]#011train-auc:0.99890#011validation-auc:0.98737
[60]#011train-auc:0.99886#011validation-auc:0.98737
[61]#011train-auc:0.99886#011validation-auc:0.98737
[62]#011train-auc:0.99886#011validation-auc:0.98737
[63]#011train-auc:0.99886#011validation-auc:0.98737
[64]#011train-auc:0.99878#011validation-auc:0.98737
[65]#011train-auc:0.99878#011validation-auc:0.98737
[66]#011train-auc:0.99898#011validation-auc:0.98737
[67]#011train-auc:0.99874#011validation-auc:0.98737
[68]#011train-auc:0.99878#011validation-auc:0.98737
[69]#011train-auc:0.99878#011validation-auc:0.98737
[70]#011train-auc:0.99878#011validation-auc:0.98737
[71]#011train-auc:0.99878#011validation-auc:0.98674
[72]#011train-auc:0.99878#011validation-auc:0.98674
[73]#011train-auc:0.99878#011validation-auc:0.98674
[74]#011train-auc:0.99878#011validation-auc:0.98674
[75]#011train-auc:0.99902#011validation-auc:0.98611
[76]#011train-auc:0.99902#011validation-auc:0.98611
[77]#011train-auc:0.99890#011validation-auc:0.98674
[78]#011train-auc:0.99878#011validation-auc:0.98674
[79]#011train-auc:0.99902#011validation-auc:0.98674
[80]#011train-auc:0.99890#011validation-auc:0.98674
[81]#011train-auc:0.99902#011validation-auc:0.98674
[82]#011train-auc:0.99914#011validation-auc:0.98674
[83]#011train-auc:0.99914#011validation-auc:0.98674
[84]#011train-auc:0.99914#011validation-auc:0.98674
[85]#011train-auc:0.99902#011validation-auc:0.98674
[86]#011train-auc:0.99902#011validation-auc:0.98674
[87]#011train-auc:0.99914#011validation-auc:0.98674
[88]#011train-auc:0.99918#011validation-auc:0.98674
[89]#011train-auc:0.99918#011validation-auc:0.98674
[90]#011train-auc:0.99906#011validation-auc:0.98674
[91]#011train-auc:0.99918#011validation-auc:0.98674
[92]#011train-auc:0.99914#011validation-auc:0.98611
[93]#011train-auc:0.99914#011validation-auc:0.98611
[94]#011train-auc:0.99914#011validation-auc:0.98674
[95]#011train-auc:0.99890#011validation-auc:0.98674
[96]#011train-auc:0.99902#011validation-auc:0.98674
[97]#011train-auc:0.99918#011validation-auc:0.98611
[98]#011train-auc:0.99906#011validation-auc:0.98674
[99]#011train-auc:0.99918#011validation-auc:0.98674
[100]#011train-auc:0.99918#011validation-auc:0.98611
[101]#011train-auc:0.99930#011validation-auc:0.98674
[102]#011train-auc:0.99918#011validation-auc:0.98611
[103]#011train-auc:0.99918#011validation-auc:0.98611
[104]#011train-auc:0.99918#011validation-auc:0.98674
[105]#011train-auc:0.99918#011validation-auc:0.98611
[106]#011train-auc:0.99918#011validation-auc:0.98674
[107]#011train-auc:0.99918#011validation-auc:0.98611
[108]#011train-auc:0.99918#011validation-auc:0.98611
```

```
[109]#011train-auc:0.99918#011validation-auc:0.98611
[110]#011train-auc:0.99918#011validation-auc:0.98611
[111]#011train-auc:0.99918#011validation-auc:0.98611
[112]#011train-auc:0.99918#011validation-auc:0.98611
[113]#011train-auc:0.99930#011validation-auc:0.98611
[114]#011train-auc:0.99918#011validation-auc:0.98611
[115]#011train-auc:0.99902#011validation-auc:0.98611
[116]#011train-auc:0.99914#011validation-auc:0.98611
[117]#011train-auc:0.99918#011validation-auc:0.98611
[118]#011train-auc:0.99902#011validation-auc:0.98611
[119]#011train-auc:0.99890#011validation-auc:0.98611
[120]#011train-auc:0.99894#011validation-auc:0.98611
[121]#011train-auc:0.99890#011validation-auc:0.98611
[122]#011train-auc:0.99906#011validation-auc:0.98548
[123]#011train-auc:0.99918#011validation-auc:0.98485
[124]#011train-auc:0.99918#011validation-auc:0.98485
[125]#011train-auc:0.99918#011validation-auc:0.98611
[126]#011train-auc:0.99918#011validation-auc:0.98611
[127]#011train-auc:0.99930#011validation-auc:0.98611
[128]#011train-auc:0.99930#011validation-auc:0.98485
[129]#011train-auc:0.99930#011validation-auc:0.98548
[130]#011train-auc:0.99930#011validation-auc:0.98548
[131]#011train-auc:0.99930#011validation-auc:0.98611
[132]#011train-auc:0.99942#011validation-auc:0.98548
[133]#011train-auc:0.99942#011validation-auc:0.98548
[134]#011train-auc:0.99942#011validation-auc:0.98611
[135]#011train-auc:0.99942#011validation-auc:0.98611
[136]#011train-auc:0.99942#011validation-auc:0.98611
[137]#011train-auc:0.99942#011validation-auc:0.98611
[138]#011train-auc:0.99942#011validation-auc:0.98674
[139]#011train-auc:0.99942#011validation-auc:0.98611
[140]#011train-auc:0.99942#011validation-auc:0.98611
[141]#011train-auc:0.99942#011validation-auc:0.98674
[142]#011train-auc:0.99930#011validation-auc:0.98674
[143]#011train-auc:0.99942#011validation-auc:0.98674
[144]#011train-auc:0.99942#011validation-auc:0.98674
[145]#011train-auc:0.99942#011validation-auc:0.98674
[146]#011train-auc:0.99942#011validation-auc:0.98674
[147]#011train-auc:0.99942#011validation-auc:0.98611
[148]#011train-auc:0.99942#011validation-auc:0.98611
[149]#011train-auc:0.99942#011validation-auc:0.98611
[150]#011train-auc:0.99942#011validation-auc:0.98611
[151]#011train-auc:0.99942#011validation-auc:0.98611
[152]#011train-auc:0.99942#011validation-auc:0.98611
[153]#011train-auc:0.99942#011validation-auc:0.98611
[154]#011train-auc:0.99942#011validation-auc:0.98611
[155]#011train-auc:0.99942#011validation-auc:0.98611
[156]#011train-auc:0.99942#011validation-auc:0.98611
[157]#011train-auc:0.99942#011validation-auc:0.98611
[158]#011train-auc:0.99942#011validation-auc:0.98674
[159]#011train-auc:0.99942#011validation-auc:0.98674
[160]#011train-auc:0.99942#011validation-auc:0.98674
[161]#011train-auc:0.99942#011validation-auc:0.98674
[162]#011train-auc:0.99942#011validation-auc:0.98674
[163]#011train-auc:0.99942#011validation-auc:0.98674
[164]#011train-auc:0.99942#011validation-auc:0.98674
```

```
[165]#011train-auc:0.99942#011validation-auc:0.98674
[166]#011train-auc:0.99942#011validation-auc:0.98674
[167]#011train-auc:0.99942#011validation-auc:0.98674
[168]#011train-auc:0.99942#011validation-auc:0.98737
[169]#011train-auc:0.99930#011validation-auc:0.98737
[170]#011train-auc:0.99942#011validation-auc:0.98737
[171]#011train-auc:0.99930#011validation-auc:0.98737
[172]#011train-auc:0.99930#011validation-auc:0.98737
[173]#011train-auc:0.99942#011validation-auc:0.98737
[174]#011train-auc:0.99930#011validation-auc:0.98737
[175]#011train-auc:0.99942#011validation-auc:0.98737
[176]#011train-auc:0.99942#011validation-auc:0.98737
[177]#011train-auc:0.99942#011validation-auc:0.98737
[178]#011train-auc:0.99942#011validation-auc:0.98737
[179]#011train-auc:0.99942#011validation-auc:0.98737
[180]#011train-auc:0.99942#011validation-auc:0.98737
[181]#011train-auc:0.99942#011validation-auc:0.98737
[182]#011train-auc:0.99942#011validation-auc:0.98737
[183]#011train-auc:0.99942#011validation-auc:0.98737
[184]#011train-auc:0.99942#011validation-auc:0.98737
[185]#011train-auc:0.99942#011validation-auc:0.98737
[186]#011train-auc:0.99942#011validation-auc:0.98737
[187]#011train-auc:0.99942#011validation-auc:0.98737
[188]#011train-auc:0.99942#011validation-auc:0.98674
[189]#011train-auc:0.99942#011validation-auc:0.98674
[190]#011train-auc:0.99942#011validation-auc:0.98674
[191]#011train-auc:0.99942#011validation-auc:0.98674
[192]#011train-auc:0.99942#011validation-auc:0.98674
[193]#011train-auc:0.99942#011validation-auc:0.98674
[194]#011train-auc:0.99942#011validation-auc:0.98674
[195]#011train-auc:0.99942#011validation-auc:0.98674
[196]#011train-auc:0.99942#011validation-auc:0.98674
[197]#011train-auc:0.99942#011validation-auc:0.98737
[198]#011train-auc:0.99942#011validation-auc:0.98737
[199]#011train-auc:0.99942#011validation-auc:0.98737

2025-09-25 10:34:51 Uploading - Uploading generated training model
2025-09-25 10:34:51 Completed - Training job completed
Training seconds: 134
Billable seconds: 134
XGBoost model artifact: s3://sagemaker-us-east-1-533267301342/sagemaker-xgboost-2025
-09-25-10-31-56-289/output/model.tar.gz
```

Evaluate XGBoost vs. baseline (on test set)

In [15]:

```
import tarfile, boto3, xgboost as xgb
from pathlib import Path
from sklearn.metrics import accuracy_score, precision_recall_fscore_support, roc_au

def parse_s3_uri(uri: str):
    assert uri.startswith("s3://")
    p = uri[5:]
    b, k = p.split("/", 1)
    return b, k
```

```
tmp_dir = Path("w4_tmp"); tmp_dir.mkdir(exist_ok=True)
bkt, key = parse_s3_uri(xgb_model_artifact)
boto3.client("s3").download_file(bkt, key, str(tmp_dir/"model.tar.gz"))
with tarfile.open(tmp_dir/"model.tar.gz") as t:
    t.extractall(tmp_dir)

# Prepare DMatrix (label stays in y_test)
dtest = xgb.DMatrix(test.drop(columns=[label_col]), label=test[label_col])
booster = xgb.Booster()
booster.load_model(str(tmp_dir/"xgboost-model"))

xgb_proba = booster.predict(dtest)
xgb_pred  = (xgb_proba >= 0.5).astype(int)

xgb_acc = accuracy_score(y_test, xgb_pred)
xgb_p, xgb_r, xgb_f1, _ = precision_recall_fscore_support(y_test, xgb_pred, average
xgb_auc = roc_auc_score(y_test, xgb_proba)

metrics_compare = {
    "baseline": baseline_metrics,  # from W4.2 (your LR baseline)
    "xgboost":  {"accuracy":xgb_acc, "precision":xgb_p, "recall":xgb_r, "f1":xgb_f1
}
metrics_compare
```

```
/tmp/ipykernel_215/4227949270.py:15: DeprecationWarning: Python 3.14 will, by defaul
t, filter extracted tar archives and reject files or modify their metadata. Use the
filter argument to control this behavior.
  t.extractall(tmp_dir)
```

Out[15]:
```
{'baseline': {'accuracy': 0.7654320987654321,
  'precision': 0.71875,
  'recall': 0.696969696969697,
  'f1': 0.7076923076923077,
  'roc_auc': 0.790719696969697},
 'xgboost': {'accuracy': 0.9876543209876543,
  'precision': 1.0,
  'recall': 0.9696969696969697,
  'f1': 0.9846153846153847,
  'roc_auc': 0.999368686868687}}
```

Deploy via Batch Transform (score Week-3 production.csv)

In [16]:
```
# Inference expects FEATURES ONLY (no label) in the SAME order as training.

from sagemaker.inputs import TransformInput

prod_df = read_csv_from_s3(f"{WEEK3_PREFIX}/production.csv")

# Same feature order as used to create training CSVs
FEATURE_COLS = [c for c in train.columns if c != label_col]
print("Feature cols count:", len(FEATURE_COLS))

prod_features = prod_df[FEATURE_COLS].copy()
bt_local = Path("w4_production_features_only.csv")
prod_features.to_csv(bt_local, index=False, header=False)
```

```
s3_bt_input = sm_session.upload_data(str(bt_local), key_prefix=f"{W4_PREFIX}/batch"

transformer = xgb_est.transformer(
    instance_count=1,
    instance_type="ml.m5.large",
    output_path=f"s3://{bucket}/{W4_PREFIX}/batch/outputs",
    accept="text/csv",
    assemble_with="Line",
)

transformer.transform(data=s3_bt_input, content_type="text/csv", split_type="Line")
transformer.wait()

batch_output_s3 = transformer.output_path
print("Batch output:", batch_output_s3)
```

```
Feature cols count: 19
INFO:sagemaker:Creating model with name: sagemaker-xgboost-2025-09-25-10-36-32-543
INFO:sagemaker:Creating transform job with name: sagemaker-xgboost-2025-09-25-10-36-
33-251
```

```
.............................../miniconda3/lib/python3.9/site-packages/sagemaker_cont
ainers/_server.py:22: UserWarning: pkg_resources is deprecated as an API. See http
s://setuptools.pypa.io/en/latest/pkg_resources.html. The pkg_resources package is sl
ated for removal as early as 2025-11-30. Refrain from using this package or pin to S
etuptools<81.
  import pkg_resources
[2025-09-25:10:41:36:INFO] No GPUs detected (normal if no gpus installed)
[2025-09-25:10:41:36:INFO] No GPUs detected (normal if no gpus installed)
[2025-09-25:10:41:36:INFO] nginx config:
worker_processes auto;
daemon off;
pid /tmp/nginx.pid;
error_log  /dev/stderr;
worker_rlimit_nofile 4096;
events {
  worker_connections 2048;
}
http {
  include /etc/nginx/mime.types;
  default_type application/octet-stream;
  access_log /dev/stdout combined;
  upstream gunicorn {
    server unix:/tmp/gunicorn.sock;
  }
  server {
    listen 8080 deferred;
    client_max_body_size 0;
    keepalive_timeout 3;
    location ~ ^/(ping|invocations|execution-parameters) {
      proxy_set_header X-Forwarded-For $proxy_add_x_forwarded_for;
      proxy_set_header Host $http_host;
      proxy_redirect off;
      proxy_read_timeout 60s;
      proxy_pass http://gunicorn;
    }
    location / {
      return 404 "{}";
    }
  }
}
[2025-09-25 10:41:36 +0000] [14] [INFO] Starting gunicorn 23.0.0
[2025-09-25 10:41:36 +0000] [14] [INFO] Listening at: unix:/tmp/gunicorn.sock (14)
[2025-09-25 10:41:36 +0000] [14] [INFO] Using worker: gevent
[2025-09-25 10:41:36 +0000] [17] [INFO] Booting worker with pid: 17
[2025-09-25 10:41:36 +0000] [18] [INFO] Booting worker with pid: 18
/miniconda3/lib/python3.9/site-packages/sagemaker_containers/_server.py:22: UserWarn
ing: pkg_resources is deprecated as an API. See https://setuptools.pypa.io/en/lates
t/pkg_resources.html. The pkg_resources package is slated for removal as early as 20
25-11-30. Refrain from using this package or pin to Setuptools<81.
  import pkg_resources
/miniconda3/lib/python3.9/site-packages/sagemaker_containers/_server.py:22: UserWarn
ing: pkg_resources is deprecated as an API. See https://setuptools.pypa.io/en/lates
t/pkg_resources.html. The pkg_resources package is slated for removal as early as 20
25-11-30. Refrain from using this package or pin to Setuptools<81.
  import pkg_resources
[2025-09-25:10:41:39:INFO] No GPUs detected (normal if no gpus installed)
```

```
[2025-09-25:10:41:39:INFO] Loading the model from /opt/ml/model/xgboost-model
[2025-09-25:10:41:39:INFO] Model objective : binary:logistic
[2025-09-25:10:41:39:INFO] No GPUs detected (normal if no gpus installed)
[2025-09-25:10:41:39:INFO] Loading the model from /opt/ml/model/xgboost-model
[2025-09-25:10:41:39:INFO] Model objective : binary:logistic
[2025-09-25:10:41:45:INFO] No GPUs detected (normal if no gpus installed)
169.254.255.130 - - [25/Sep/2025:10:41:45 +0000] "GET /ping HTTP/1.1" 200 0 "-" "Go-
http-client/1.1"
[2025-09-25:10:41:45:INFO] No GPUs detected (normal if no gpus installed)
169.254.255.130 - - [25/Sep/2025:10:41:45 +0000] "GET /execution-parameters HTTP/1.
1" 200 84 "-" "Go-http-client/1.1"
[2025-09-25:10:41:45:INFO] Determined delimiter of CSV input is ','
/miniconda3/lib/python3.9/site-packages/xgboost/core.py:122: UserWarning: ntree_limi
t is deprecated, use `iteration_range` or model slicing instead.
  warnings.warn(
169.254.255.130 - - [25/Sep/2025:10:41:45 +0000] "POST /invocations HTTP/1.1" 200 65
83 "-" "Go-http-client/1.1"
[2025-09-25:10:41:45:INFO] No GPUs detected (normal if no gpus installed)
169.254.255.130 - - [25/Sep/2025:10:41:45 +0000] "GET /ping HTTP/1.1" 200 0 "-" "Go-
http-client/1.1"
[2025-09-25:10:41:45:INFO] No GPUs detected (normal if no gpus installed)
169.254.255.130 - - [25/Sep/2025:10:41:45 +0000] "GET /execution-parameters HTTP/1.
1" 200 84 "-" "Go-http-client/1.1"
[2025-09-25:10:41:45:INFO] Determined delimiter of CSV input is ','
/miniconda3/lib/python3.9/site-packages/xgboost/core.py:122: UserWarning: ntree_limi
t is deprecated, use `iteration_range` or model slicing instead.
  warnings.warn(
169.254.255.130 - - [25/Sep/2025:10:41:45 +0000] "POST /invocations HTTP/1.1" 200 65
83 "-" "Go-http-client/1.1"
2025-09-25T10:41:45.426:[sagemaker logs]: MaxConcurrentTransforms=2, MaxPayloadInMB=
6, BatchStrategy=MULTI_RECORD

/miniconda3/lib/python3.9/site-packages/sagemaker_containers/_server.py:22: UserWarn
ing: pkg_resources is deprecated as an API. See https://setuptools.pypa.io/en/lates
t/pkg_resources.html. The pkg_resources package is slated for removal as early as 20
25-11-30. Refrain from using this package or pin to Setuptools<81.
  import pkg_resources
[2025-09-25:10:41:36:INFO] No GPUs detected (normal if no gpus installed)
[2025-09-25:10:41:36:INFO] No GPUs detected (normal if no gpus installed)
[2025-09-25:10:41:36:INFO] nginx config:
worker_processes auto;
daemon off;
pid /tmp/nginx.pid;
error_log  /dev/stderr;
worker_rlimit_nofile 4096;
events {
  worker_connections 2048;
/miniconda3/lib/python3.9/site-packages/sagemaker_containers/_server.py:22: UserWarn
ing: pkg_resources is deprecated as an API. See https://setuptools.pypa.io/en/lates
t/pkg_resources.html. The pkg_resources package is slated for removal as early as 20
25-11-30. Refrain from using this package or pin to Setuptools<81.
  import pkg_resources
[2025-09-25:10:41:36:INFO] No GPUs detected (normal if no gpus installed)
[2025-09-25:10:41:36:INFO] No GPUs detected (normal if no gpus installed)
[2025-09-25:10:41:36:INFO] nginx config:
worker_processes auto;
```

```
daemon off;
pid /tmp/nginx.pid;
error_log  /dev/stderr;
worker_rlimit_nofile 4096;
events {
  worker_connections 2048;
}
http {
  include /etc/nginx/mime.types;
  default_type application/octet-stream;
  access_log /dev/stdout combined;
  upstream gunicorn {
    server unix:/tmp/gunicorn.sock;
  }
  server {
    listen 8080 deferred;
    client_max_body_size 0;
    keepalive_timeout 3;
    location ~ ^/(ping|invocations|execution-parameters) {
      proxy_set_header X-Forwarded-For $proxy_add_x_forwarded_for;
      proxy_set_header Host $http_host;
      proxy_redirect off;
      proxy_read_timeout 60s;
      proxy_pass http://gunicorn;
    }
    location / {
      return 404 "{}";
    }
  }
}
[2025-09-25 10:41:36 +0000] [14] [INFO] Starting gunicorn 23.0.0
[2025-09-25 10:41:36 +0000] [14] [INFO] Listening at: unix:/tmp/gunicorn.sock (14)
[2025-09-25 10:41:36 +0000] [14] [INFO] Using worker: gevent
[2025-09-25 10:41:36 +0000] [17] [INFO] Booting worker with pid: 17
[2025-09-25 10:41:36 +0000] [18] [INFO] Booting worker with pid: 18
/miniconda3/lib/python3.9/site-packages/sagemaker_containers/_server.py:22: UserWarn
ing: pkg_resources is deprecated as an API. See https://setuptools.pypa.io/en/lates
t/pkg_resources.html. The pkg_resources package is slated for removal as early as 20
25-11-30. Refrain from using this package or pin to Setuptools<81.
  import pkg_resources
/miniconda3/lib/python3.9/site-packages/sagemaker_containers/_server.py:22: UserWarn
ing: pkg_resources is deprecated as an API. See https://setuptools.pypa.io/en/lates
t/pkg_resources.html. The pkg_resources package is slated for removal as early as 20
25-11-30. Refrain from using this package or pin to Setuptools<81.
  import pkg_resources
}
http {
  include /etc/nginx/mime.types;
  default_type application/octet-stream;
  access_log /dev/stdout combined;
  upstream gunicorn {
    server unix:/tmp/gunicorn.sock;
  }
  server {
    listen 8080 deferred;
    client_max_body_size 0;
```

```
    keepalive_timeout 3;
    location ~ ^/(ping|invocations|execution-parameters) {
      proxy_set_header X-Forwarded-For $proxy_add_x_forwarded_for;
      proxy_set_header Host $http_host;
      proxy_redirect off;
      proxy_read_timeout 60s;
      proxy_pass http://gunicorn;
    }
    location / {
      return 404 "{}";
    }
  }
}
[2025-09-25 10:41:36 +0000] [14] [INFO] Starting gunicorn 23.0.0
[2025-09-25 10:41:36 +0000] [14] [INFO] Listening at: unix:/tmp/gunicorn.sock (14)
[2025-09-25 10:41:36 +0000] [14] [INFO] Using worker: gevent
[2025-09-25 10:41:36 +0000] [17] [INFO] Booting worker with pid: 17
[2025-09-25 10:41:36 +0000] [18] [INFO] Booting worker with pid: 18
/miniconda3/lib/python3.9/site-packages/sagemaker_containers/_server.py:22: UserWarn
ing: pkg_resources is deprecated as an API. See https://setuptools.pypa.io/en/lates
t/pkg_resources.html. The pkg_resources package is slated for removal as early as 20
25-11-30. Refrain from using this package or pin to Setuptools<81.
  import pkg_resources
/miniconda3/lib/python3.9/site-packages/sagemaker_containers/_server.py:22: UserWarn
ing: pkg_resources is deprecated as an API. See https://setuptools.pypa.io/en/lates
t/pkg_resources.html. The pkg_resources package is slated for removal as early as 20
25-11-30. Refrain from using this package or pin to Setuptools<81.
  import pkg_resources
[2025-09-25:10:41:39:INFO] No GPUs detected (normal if no gpus installed)
[2025-09-25:10:41:39:INFO] Loading the model from /opt/ml/model/xgboost-model
[2025-09-25:10:41:39:INFO] Model objective : binary:logistic
[2025-09-25:10:41:39:INFO] No GPUs detected (normal if no gpus installed)
[2025-09-25:10:41:39:INFO] Loading the model from /opt/ml/model/xgboost-model
[2025-09-25:10:41:39:INFO] Model objective : binary:logistic
[2025-09-25:10:41:39:INFO] No GPUs detected (normal if no gpus installed)
[2025-09-25:10:41:39:INFO] Loading the model from /opt/ml/model/xgboost-model
[2025-09-25:10:41:39:INFO] Model objective : binary:logistic
[2025-09-25:10:41:39:INFO] No GPUs detected (normal if no gpus installed)
[2025-09-25:10:41:39:INFO] Loading the model from /opt/ml/model/xgboost-model
[2025-09-25:10:41:39:INFO] Model objective : binary:logistic
[2025-09-25:10:41:45:INFO] No GPUs detected (normal if no gpus installed)
169.254.255.130 - - [25/Sep/2025:10:41:45 +0000] "GET /ping HTTP/1.1" 200 0 "-" "Go-
http-client/1.1"
[2025-09-25:10:41:45:INFO] No GPUs detected (normal if no gpus installed)
169.254.255.130 - - [25/Sep/2025:10:41:45 +0000] "GET /execution-parameters HTTP/1.
1" 200 84 "-" "Go-http-client/1.1"
[2025-09-25:10:41:45:INFO] Determined delimiter of CSV input is ','
/miniconda3/lib/python3.9/site-packages/xgboost/core.py:122: UserWarning: ntree_limi
t is deprecated, use `iteration_range` or model slicing instead.
  warnings.warn(
169.254.255.130 - - [25/Sep/2025:10:41:45 +0000] "POST /invocations HTTP/1.1" 200 65
83 "-" "Go-http-client/1.1"
[2025-09-25:10:41:45:INFO] No GPUs detected (normal if no gpus installed)
169.254.255.130 - - [25/Sep/2025:10:41:45 +0000] "GET /ping HTTP/1.1" 200 0 "-" "Go-
http-client/1.1"
[2025-09-25:10:41:45:INFO] No GPUs detected (normal if no gpus installed)
```

```
169.254.255.130 - - [25/Sep/2025:10:41:45 +0000] "GET /execution-parameters HTTP/1.
1" 200 84 "-" "Go-http-client/1.1"
[2025-09-25:10:41:45:INFO] Determined delimiter of CSV input is ','
/miniconda3/lib/python3.9/site-packages/xgboost/core.py:122: UserWarning: ntree_limi
t is deprecated, use `iteration_range` or model slicing instead.
  warnings.warn(
169.254.255.130 - - [25/Sep/2025:10:41:45 +0000] "POST /invocations HTTP/1.1" 200 65
83 "-" "Go-http-client/1.1"
2025-09-25T10:41:45.426:[sagemaker logs]: MaxConcurrentTransforms=2, MaxPayloadInMB=
6, BatchStrategy=MULTI_RECORD
Batch output: s3://sagemaker-us-east-1-533267301342/aai540/maternal-risk/week4/20250
925-102857/batch/outputs
```

Artifacts, design-doc snippet, tracker update

In [17]:

```
import json, boto3, matplotlib.pyplot as plt
from sklearn.metrics import confusion_matrix

def plot_cm(cm, title, path):
    plt.figure()
    plt.imshow(cm, interpolation='nearest'); plt.title(title); plt.colorbar()
    plt.xlabel("Predicted"); plt.ylabel("Actual"); plt.tight_layout(); plt.savefig(

# Confusion matrices
bm_pred = (bm_clf.predict_proba(test[["Age","SystolicBP"]])[:,1] >= 0.5).astype(int
bm_cm   = confusion_matrix(y_test, bm_pred)
xgb_cm  = confusion_matrix(y_test, xgb_pred)

# Save locally
from pathlib import Path
art_dir = Path("w4_artifacts"); art_dir.mkdir(exist_ok=True)
plot_cm(bm_cm,  "Baseline CM", art_dir/"baseline_cm.png")
plot_cm(xgb_cm, "XGBoost CM",  art_dir/"xgb_cm.png")

metrics_compare = {
    "baseline": baseline_metrics,
    "xgboost":  {
        "accuracy": float(metrics_compare["xgboost"]["accuracy"]),
        "precision": float(metrics_compare["xgboost"]["precision"]),
        "recall": float(metrics_compare["xgboost"]["recall"]),
        "f1": float(metrics_compare["xgboost"]["f1"]),
        "roc_auc": float(metrics_compare["xgboost"]["roc_auc"]),
    },
}
with open(art_dir/"metrics_compare.json","w") as f:
    json.dump(metrics_compare, f, indent=2)

# Upload artifacts
s3c = boto3.client("s3")
def up(local, key):
    s3c.upload_file(str(local), bucket, f"{W4_PREFIX}/{key}")
    return f"s3://{bucket}/{W4_PREFIX}/{key}"

metrics_s3 = up(art_dir/"metrics_compare.json", "metrics_compare.json")
bm_cm_s3   = up(art_dir/"baseline_cm.png",      "baseline_cm.png")
```

```
xgb_cm_s3  = up(art_dir/"xgb_cm.png",           "xgb_cm.png")

# Design-Doc snippet (paste this block into your ML Design Document)
design_doc_snippet = f"""
### Week 4 Findings — Model Development & Deployment

**Benchmark (LogReg on Age + SystolicBP)**
Acc: {baseline_metrics['accuracy']:.3f} | Prec: {baseline_metrics['precision']:.3f}

**XGBoost (full features)**
Acc: {metrics_compare['xgboost']['accuracy']:.3f} | Prec: {metrics_compare['xgboost

**Artifacts**
- Metrics JSON: {metrics_s3}
- Baseline CM:  {bm_cm_s3}
- XGBoost CM:   {xgb_cm_s3}
- XGBoost Model Artifact: {xgb_model_artifact}
- Batch Transform Output: {batch_output_s3}
"""
print(design_doc_snippet)

# Tracker (JSON + Markdown) → S3
w4_tracker_dir = Path("w4_tracker"); w4_tracker_dir.mkdir(exist_ok=True)
tracker_w4 = {
    "week": "4",
    "run_id": RUN_ID,
    "week3_prefix": f"s3://{bucket}/{WEEK3_PREFIX}",
    "week4_prefix": f"s3://{bucket}/{W4_PREFIX}",
    "benchmark": baseline_metrics,
    "xgboost": metrics_compare["xgboost"],
    "artifacts": {
        "metrics_json": metrics_s3,
        "baseline_cm": bm_cm_s3,
        "xgb_cm": xgb_cm_s3,
        "model_artifact": xgb_model_artifact,
        "batch_output": batch_output_s3
    }
}
with open(w4_tracker_dir/"team_tracker_update_week4.json","w") as f:
    json.dump(tracker_w4, f, indent=2)

md = f"""# Week 4 Tracker – Maternal Health Risk (RUN: {RUN_ID})

**Week-3 prefix:** s3://{bucket}/{WEEK3_PREFIX}
**Week-4 prefix:** s3://{bucket}/{W4_PREFIX}

## Benchmark (LogReg on Age + SystolicBP)
Acc: {baseline_metrics['accuracy']:.3f} | Prec: {baseline_metrics['precision']:.3f}

## XGBoost (full features)
Acc: {metrics_compare['xgboost']['accuracy']:.3f} | Prec: {metrics_compare['xgboost

## Artifacts
- Metrics JSON: {metrics_s3}
- Baseline CM:  {bm_cm_s3}
- XGBoost CM:   {xgb_cm_s3}
```

```
- Model:        {xgb_model_artifact}
- Batch Output: {batch_output_s3}
"""
with open(w4_tracker_dir/"team_tracker_update_week4.md","w") as f:
    f.write(md)

s3c.upload_file(str(w4_tracker_dir/"team_tracker_update_week4.json"), bucket, f"{W4
s3c.upload_file(str(w4_tracker_dir/"team_tracker_update_week4.md"),   bucket, f"{W4

print("Week-4 tracker written & uploaded →", f"s3://{bucket}/{W4_PREFIX}/team_track
```

```
### Week 4 Findings — Model Development & Deployment

**Benchmark (LogReg on Age + SystolicBP)**
Acc: 0.765 | Prec: 0.719 | Rec: 0.697 | F1: 0.708 | AUC: 0.791

**XGBoost (full features)**
Acc: 0.988 | Prec: 1.000 | Rec: 0.970 | F1: 0.985 | AUC: 0.999

**Artifacts**
- Metrics JSON: s3://sagemaker-us-east-1-533267301342/aai540/maternal-risk/week4/202
50925-102857/metrics_compare.json
- Baseline CM:  s3://sagemaker-us-east-1-533267301342/aai540/maternal-risk/week4/202
50925-102857/baseline_cm.png
- XGBoost CM:   s3://sagemaker-us-east-1-533267301342/aai540/maternal-risk/week4/202
50925-102857/xgb_cm.png
- XGBoost Model Artifact: s3://sagemaker-us-east-1-533267301342/sagemaker-xgboost-20
25-09-25-10-31-56-289/output/model.tar.gz
- Batch Transform Output: s3://sagemaker-us-east-1-533267301342/aai540/maternal-ris
k/week4/20250925-102857/batch/outputs

Week-4 tracker written & uploaded → s3://sagemaker-us-east-1-533267301342/aai540/mat
ernal-risk/week4/20250925-102857/team_tracker_update_week4.*
```

In [ ]: