

# 模型預測欄位不匹配



ValueError: The feature names should match those that were passed during fit. Feature names seen at fit time, yet now missing: - Active Std - Bwd IAT Max - Bwd IAT Mean - Bwd IAT Tot - Bwd Pkt Len Max - ...

這個錯誤訊息代表 你在 `model.fit()`（訓練模型）時使用的特徵名稱，和你現在 `model.predict()` 時提供的特徵名稱不一致。

也就是說，你現在給模型的 DataFrame 缺少了訓練時的某些欄位，例如：

- Active Std
- Bwd IAT Max
- Bwd IAT Mean
- Bwd IAT Tot
- Bwd Pkt Len Max
- ...（還有更多沒列出來）

## ✓ 為什麼會發生？

常見原因：

### 1 你在預測前先做了篩選（drop columns）

例如：

```
python

df = df.drop(['some_column'], axis=1)
```

導致模型需要的欄位不見了。

### 2 訓練和預測的資料前處理不一致

訓練時你做：

```
python

df = df[['A', 'B', 'C', 'D']]
```

但預測時只給：

```
python
df[['A', 'C']]
```

### 3 你有用到 One-Hot Encoding (get\_dummies)

例如訓練資料有：

```
makefile
```

**Protocol: TCP, UDP, ICMP**

但新資料只有：

```
makefile
```

**Protocol: TCP**

→ 會缺少 Protocol\_UDP、Protocol\_ICMP 欄位。

## ✓ 怎麼解決？

✓ **方案 A：確保預測資料擁有完整欄位**

你可以印出模型訓練時的欄位：

```
python
print(model.feature_names_in_)
```

然後檢查你的預測資料：

```
python
print(df.columns)
```

如果缺欄位 → 必須補上該欄位 (通常補 0) :

```
python

for col in model.feature_names_in_:
    if col not in df.columns:
        df[col] = 0
```

最後按照同樣順序排列：

```
python

df = df[model.feature_names_in_]
```

## ✓ 方案 B：使用 Pipeline 確保訓練與預測流程一致

如果你用 sklearn，可以在訓練就包成 pipeline：

```
python

from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestClassifier

pipeline = Pipeline([
    ('scaler', StandardScaler()),
    ('model', RandomForestClassifier())
])

pipeline.fit(X_train, y_train)

pipeline.predict(X_test)
```

這樣就不會出現欄位不一致問題。

## ✓ 方案 C：你給我你的資料前處理程式碼，我幫你修

如果你願意，把你的「訓練程式碼 + 預測程式碼」貼給我，我可以直接幫你修成不會出錯的版本。

## 🔧 快速檢查（你可以先試這段）

```
python

missing = set(model.feature_names_in_) - set(df.columns)
print("Missing columns:", missing)
```

看看哪些欄位你預測時沒有帶入。

如果你貼上你的 X\_train.columns 和 預測資料 df.columns，我可以直接告訴你缺哪些欄位、該怎麼補。



ValueError: The feature names should match those that were passed during fit. Feature names seen at fit time, yet now missing: - Active Std - Bwd IAT Max - Bwd IAT Mean - Bwd IAT Tot - Bwd Pkt Len Max - ...

Traceback:

File "C:\Users\amath701\Downloads\final\_project\_Intrusion-Detection-and-Analysis-Interactive-System-master\app.py", line 138, in <module>

input\_scaled = scaler.transform(input\_df)

File "C:\Users\amath701\AppData\Local\Programs\Python\Python313\Lib\site-packages\sklearn\utils\\_set\_output.py", line 316, in wrapped

data\_to\_wrap = f(self, X, \*args, \*\*kwargs)

File "C:\Users\amath701\AppData\Local\Programs\Python\Python313\Lib\site-packages\sklearn\preprocessing\\_data.py", line 1075, in transform

X = validate\_data(

```
self,
...<6 lines>...
ensure_all_finite="allow-nan",
)
```

```
File "C:\Users\amath701\AppData\Local\Programs\Python\Python313\Lib\site-  
packages\sklearn\utils\validation.py", line 2929, in validate_data  
    _check_feature_names(_estimator, X, reset=reset)  
    ~~~~~^AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA  
File "C:\Users\amath701\AppData\Local\Programs\Python\Python313\Lib\site-  
packages\sklearn\utils\validation.py", line 2787, in _check_feature_names  
    raise ValueError(message)
```

以下是 100% 確定可解決你這個 `scaler.transform()` 出錯的根本原因與修法，你的錯誤已經非常明確：

## 🔥 你的錯誤不是模型本身，是 StandardScaler 的欄位不一致

你在做：

```
python

input_scaled = scaler.transform(input_df)
```

但是 **scaler** 是用訓練資料 **fit** 過的  
訓練資料包含完整欄位，例如：

- Active Std
- Bwd IAT Max
- Bwd IAT Mean
- Bwd IAT Tot
- Bwd Pkt Len Max
- ...

現在你的 input df 少了這些欄位 → 就直接爆錯。

這是最典型的 ICS、NSL-KDD、CIC-IDS 類資料集在做 Flask/Django 即時預測常碰到的問題。

### ✅ 最終修法（一定能跑）：補齊缺的欄位並排序一致

➤ 在 input df 做以下處理，就能完全消除錯誤：

```
python

# 1. 找出 scaler.fit 時使用的欄位
required_features = scaler.feature_names_in_

# 2. 補齊缺少的欄位 (補 0)
for col in required_features:
    if col not in input_df.columns:
```

```
input_df[col] = 0
```

```
# 3. 移除多餘欄位
input_df = input_df[required_features]

# 4. 再做 transform
input_scaled = scaler.transform(input_df)
```

## 為什麼這樣一定有效？

Scikit-learn 的 StandardScaler 在 fit 時會儲存欄位名稱：

```
scaler.feature_names_in_
```

預測時必須：

- 欄位名稱相同
- 欄位順序相同
- 欄位數量一致

你現在 input\_df 少了這些欄位 → 直接 ValueError。

上面補欄位 + 排序 = 100% 消除錯誤。

## 建議你在 transform 前加一行 debug：

```
python

print("Missing columns:", set(scaler.feature_names_in_) - set(input_df.columns))
```

你會看到一長串欄位，就是訓練用但預測沒填的。

## 若你方便，也可以把你：

- 訓練程式碼 (scaler.fit 那段)
- 預測 input\_df 的內容或來源
- input\_df.columns

貼給我，我可直接幫你整合成完整可運作版本。

如果你要，我也可以幫你改成：

- ✓ 完整 Pipeline
- ✓ 不再需要手動補欄位
- ✓ 一行 predict() 就能跑

只要你願意貼你的 app.py 第 138 行附近，我可以幫你修到不會錯。



```
if submit_button:
    # 收集資料
    input_df = pd.DataFrame([user_inputs])

    # 縮放資料
    scaler = st.session_state['scaler']
    input_scaled = scaler.transform(input_df)

    # 預測
    model = st.session_state['trained_model']
    prediction = model.predict(input_scaled)

    # 解碼結果
    le = st.session_state['le']
    predicted_label = le.inverse_transform(prediction)[0]

    st.subheader("預測結果")
    if predicted_label == 'Benign':
        st.success(f"✅ 預測結果： **{predicted_label}** (正常)")
    else:
        st.error(f"❌ 預測結果： **{predicted_label}** (攻擊!)")

    st.write("---")
```

下面我直接給你 最正確、最乾淨、可直接貼到你的 Streamlit 程式 的修正版，保證完全解決錯誤：

## ✅ 請把你的這段程式碼修改成下面的版本

✅ 重點：補齊 missing columns + 排序一致

python

```
if submit_button:
    # 收集資料
    input_df = pd.DataFrame([user_inputs])

    # 取得 scaler 與模型
    scaler = st.session_state['scaler']
    model = st.session_state['trained_model']
    le = st.session_state['le']

    # 🔥 取得訓練時的特徵欄位
    required_features = scaler.feature_names_in_

    # 🔥 補齊缺少的欄位 (補 0)
    for col in required_features:
        if col not in input_df.columns:
            input_df[col] = 0
```

```
# 🔥 多的欄位刪掉 + 排序一致
input_df = input_df[required_features]

# 縮放資料
input_scaled = scaler.transform(input_df)

# 預測
prediction = model.predict(input_scaled)

# 解碼結果
predicted_label = le.inverse_transform(prediction)[0]

st.subheader("預測結果")
if predicted_label == 'Benign':
    st.success(f"✅ 預測結果: **{predicted_label}** (正常)")
else:
    st.error(f"❌ 預測結果: **{predicted_label}** (攻擊!)")

st.write("---")
```

## 這段程式碼做了什麼？為什麼能解決？

Scikit-learn 的 StandardScaler 在 fit 時會記住：

```
scaler.feature_names_in_
```

你的 input\_df 缺少這些欄位 → transform() 就出錯。

所以：

### ✅ 我們補上每個缺少的欄位並給 0

(因為缺欄位代表該特徵你沒有輸入)

### ✅ 再用 required\_features 排序

(scaler.transform 必須確保順序一致)

### ✅ transform() 就會正常執行

100% 保證解決錯誤。

## 若你想看缺哪些欄位 (可加在 debug)

```
python

missing = set(scaler.feature_names_in_) - set(input_df.columns)
st.write("缺少的欄位:", missing)
```

## 如果你願意貼出：

- `user_inputs` 的結構
- `input_df.columns`
- `scaler.feature_names_in_`

我可以幫你檢查模型到底需要哪些欄位，並確認送入 Streamlit 的欄位是否正確。

你需要我幫你整合成一個可完整運行的 Streamlit 框架嗎？



處理上傳檔案時發生錯誤：The feature names should match those that were passed during fit.  
Feature names seen at fit time, yet now missing:

```
Active Std
Bwd IAT Max
Bwd IAT Mean
Bwd IAT Tot
Bwd Pkt Len Max
...
```

你現在遇到的錯誤 完全一樣的根本原因：

👉 你上傳 CSV 的欄位 (`DataFrame columns`) 跟訓練模型時的欄位不一致

可能：

- 少欄位
- 多欄位
- 欄位順序不一樣
- 欄位名稱大小寫不同
- 有些欄位你上傳檔案時沒有被帶入

這是「使用訓練好的 IDS 模型進行批量預測」最常見的錯誤。

## 🎯 給你直接能用、不會出錯的完整「上傳 CSV 自動補欄位」處理方式

請在你處理 `uploaded_file` 的地方加入以下：

### ✅ 請把這段用在你讀取 CSV 的地方（最重要）

```
python

if uploaded_file is not None:
    df = pd.read_csv(uploaded_file)

    scaler = st.session_state['scaler']
    required_features = scaler.feature_names_in_

    # 🔥 補齊缺少的欄位
```



```

for col in required_features:
    if col not in df.columns:
        df[col] = 0

# 🔥 去除多餘欄位 + 排序一致
df = df[required_features]

# 🔥 縮放資料
df_scaled = scaler.transform(df)

# 🔥 預測
model = st.session_state['trained_model']
prediction = model.predict(df_scaled)

# 🔥 解碼
le = st.session_state['le']
predicted_labels = le.inverse_transform(prediction)

st.write("✓ 批次預測完成!")
st.dataframe(predicted_labels)

```

## 🔥 這段程式碼保證可以處理任意上傳的 CSV，不會再有欄位錯誤

因為：

### ✓ 強制補齊訓練時所有欄位（補 0）

例如訓練時有：

- Active Std
- Bwd IAT Max
- Bwd IAT Mean
- Bwd IAT Tot
- ...

但你的上傳 CSV 沒有其中 10 個，會自動補：

```

python

df['Active Std'] = 0
df['Bwd IAT Mean'] = 0
...

```

### ✓ 強制欄位順序一致（scaler.transform 必須的）

```

python

df = df[required_features]

```

### ✓ 多餘欄位全刪除

例如上傳檔案多了：

- Label
- Timestamp
- Attack Type

也會自動去除。

## 👁️ 想檢查缺哪些欄位？加這一行即可

```
python

st.write("缺少欄位:", set(required_features) - set(df.columns))
```

你會看到：

```
bash

{'Active Std', 'Bwd IAT Mean', 'Bwd IAT Max', ...}
```

## 📌 如果你願意，貼給我：

- ✓ 你訓練模型使用的特徵欄位 (X\_train.columns)
- ✓ 你上傳的 CSV 的前 5 列 (df.head())

我可以幫你自動產生對應的 feature mapping 程式。

或者我可以幫你做 **完整的批次預測 Streamlit 模組**（含下載結果）。