📊 Project 6 : Model Explainability with SHA

In this notebook, we apply **SHAP (SHapley Additive exPlanations)** to interpret machine learnin models predicting employee attrition.

Goals:

- Understand which features drive attrition globally across employees
- Explain individual predictions for specific employees
- Provide visuals that HR leaders can use to trust model decisions

1️⃣ Imports

2️⃣ Load Data & Model

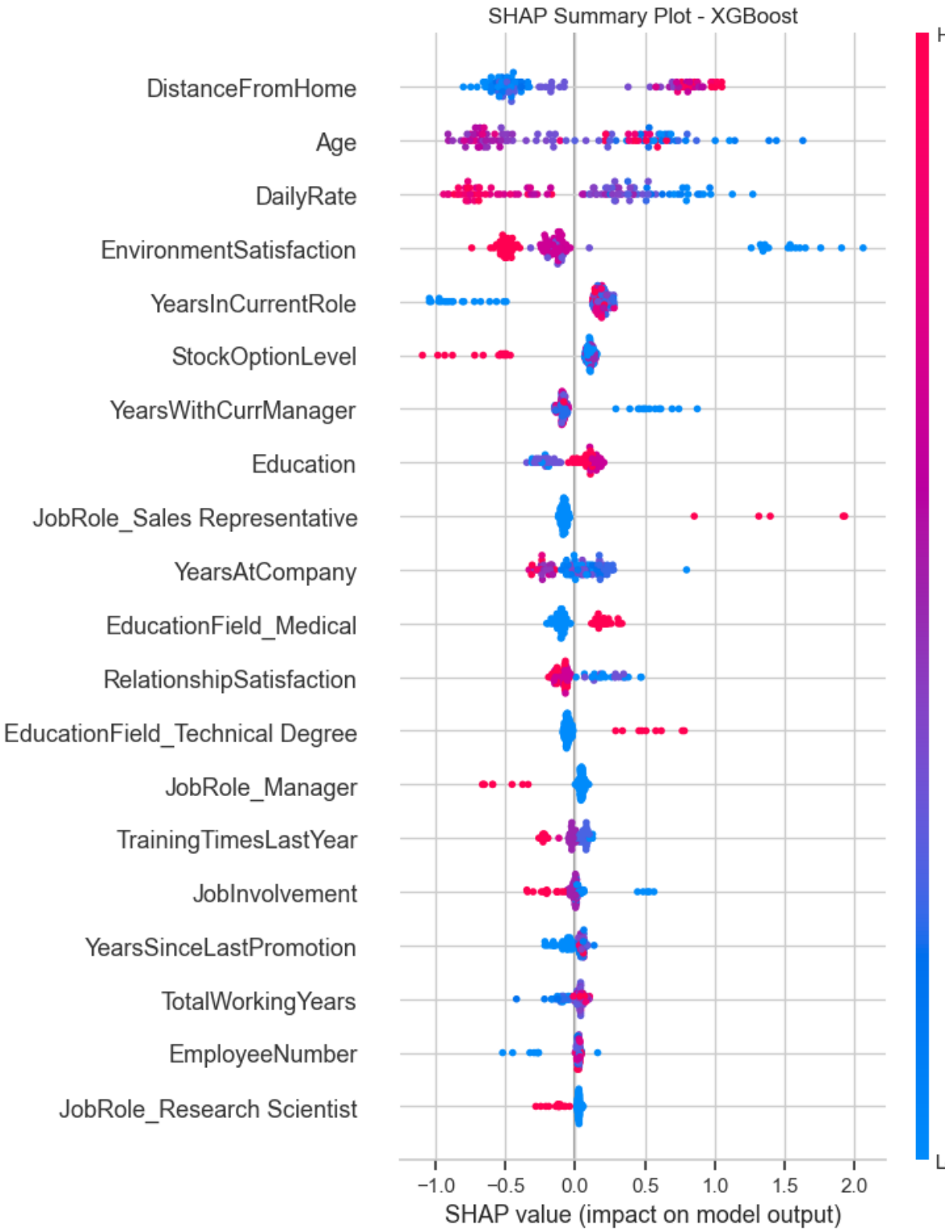
```
Dataset shape: (1470, 47)
```

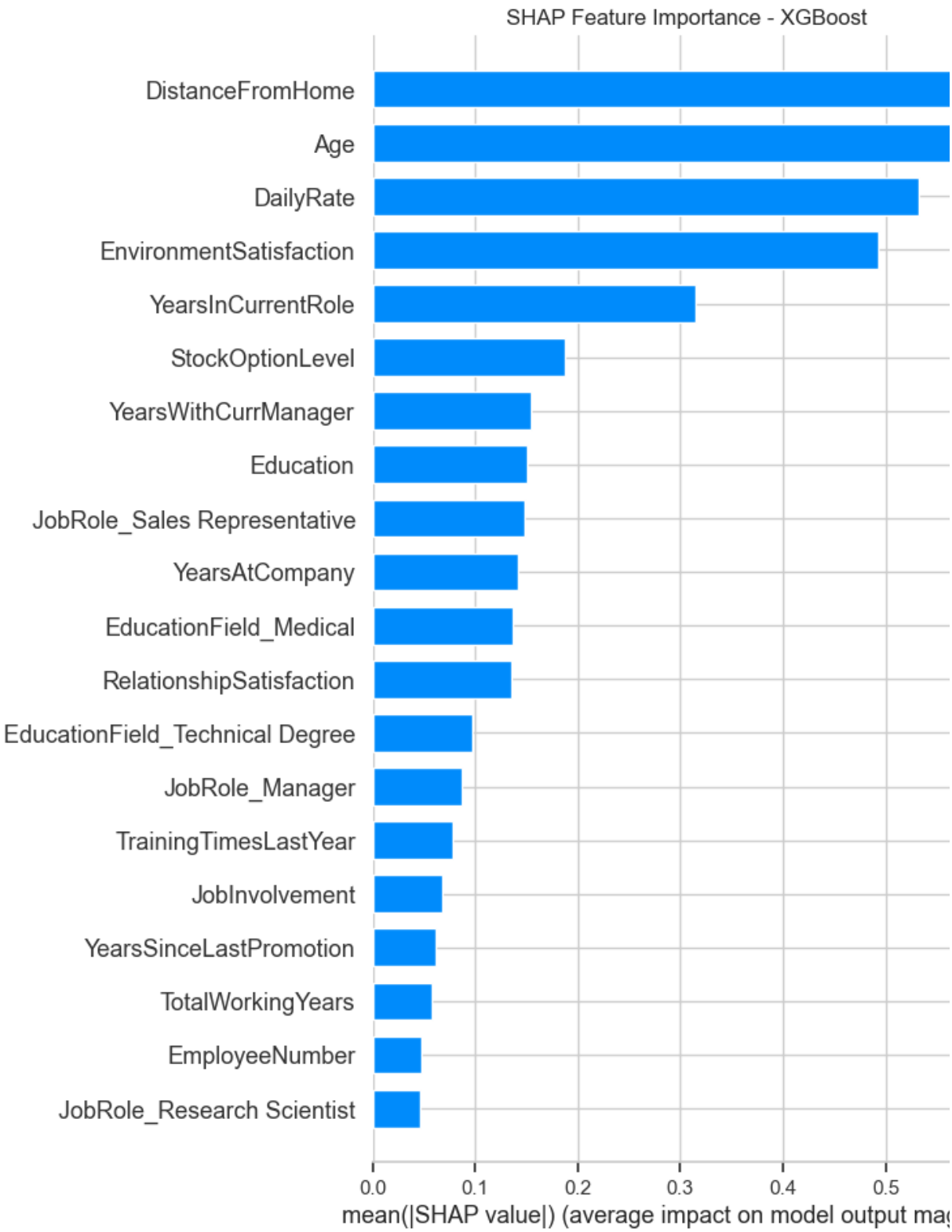
3️⃣ Subsample for SHAP (⚡ Stability Fix)

```
SHAP values computed on sample: (100, 47)
```

🌍 Global Interpretability

Visuals showing which features have the strongest impact on attrition risk across employees.

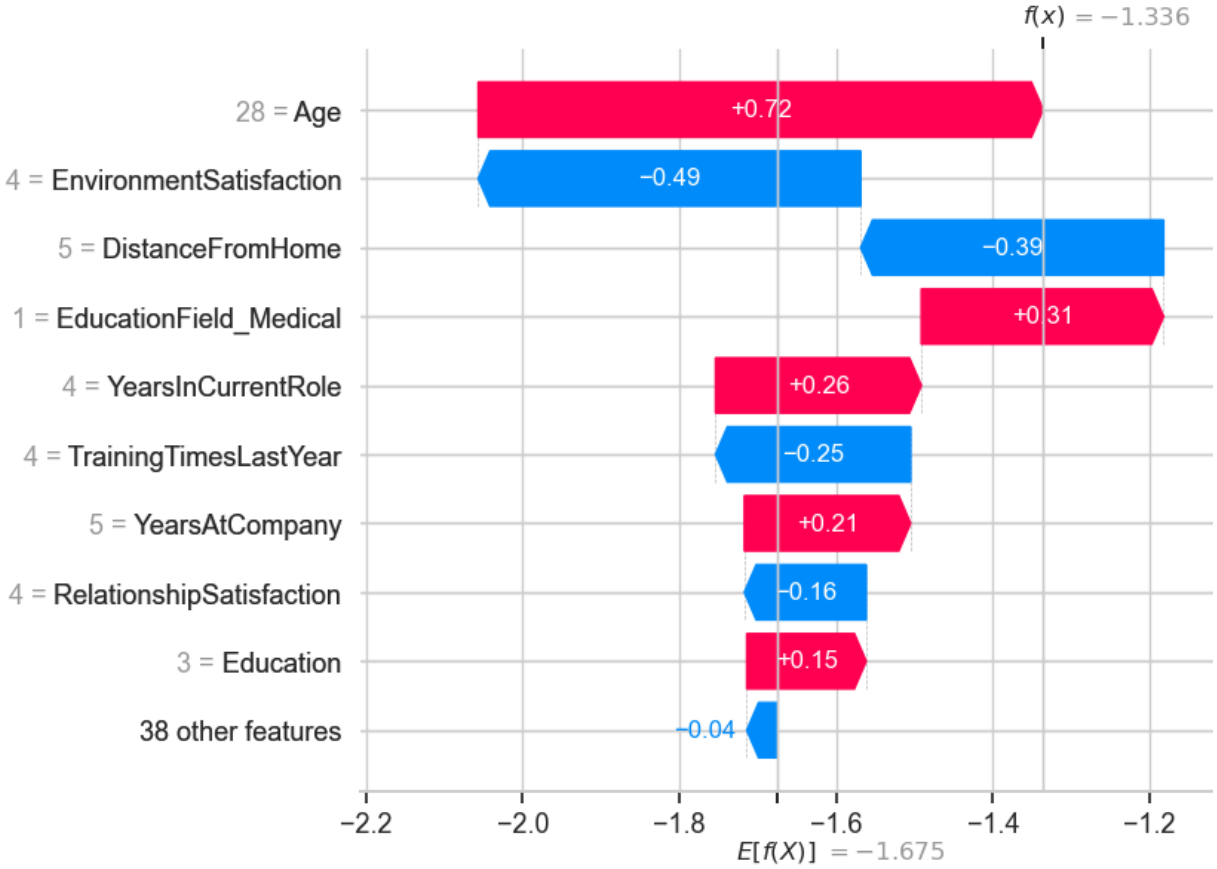

SHAP Summary Plot - XGBoost
DistanceFromHome
Age
DailyRate
EnvironmentSatisfaction
YearsInCurrentRole
StockOptionLevel
YearsWithCurrManager
Education
JobRole_Sales Representative
YearsAtCompany
EducationField_Medical
RelationshipSatisfaction
EducationField_Technical Degree
JobRole_Manager
TrainingTimesLastYear
JobInvolvement
YearsSinceLastPromotion
TotalWorkingYears
EmployeeNumber
JobRole_Research Scientist
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
SHAP value (impact on model output)
H
L

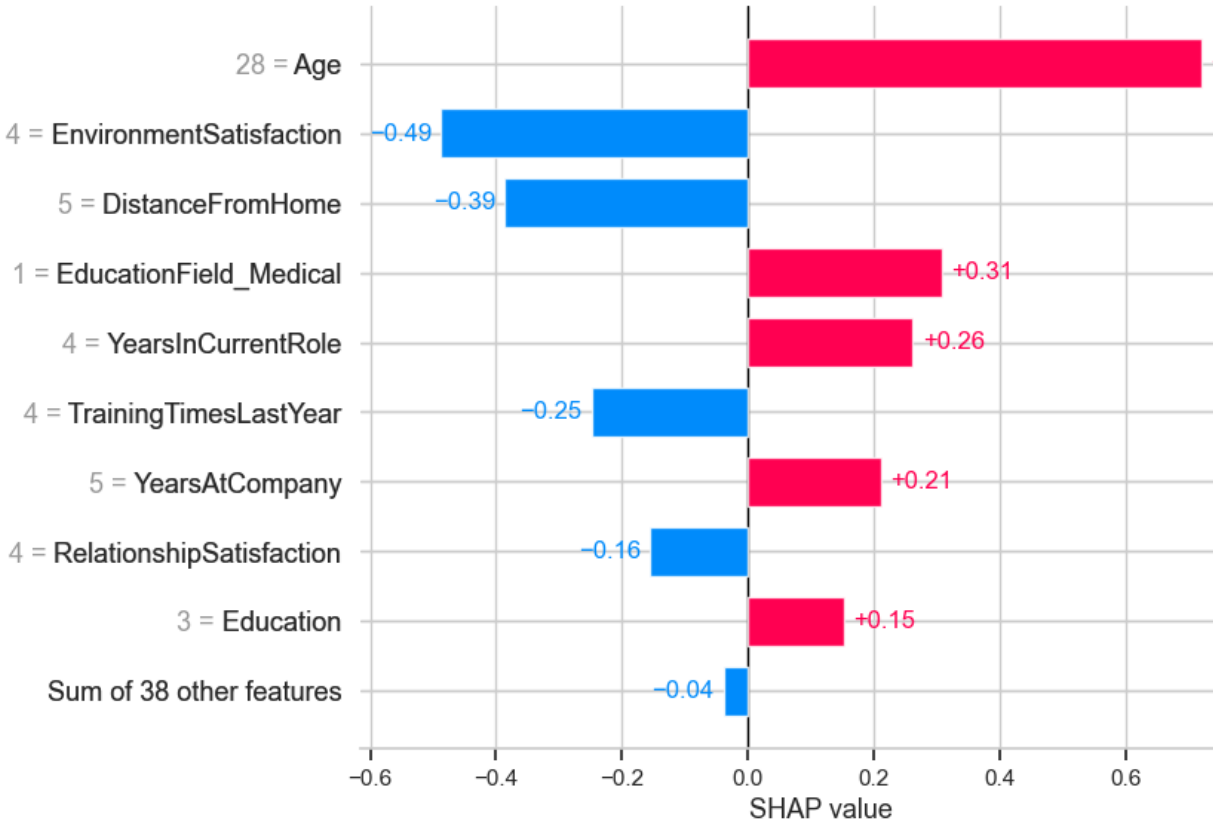


👤 Local Interpretability (Waterfall + Bar)

For a specific employee, SHAP can explain how each feature contributed to their attrition risk. We'll show two visuals:

- **Waterfall Plot** → Directional push of each feature
- **Bar Plot** → Top features ranked by contribution

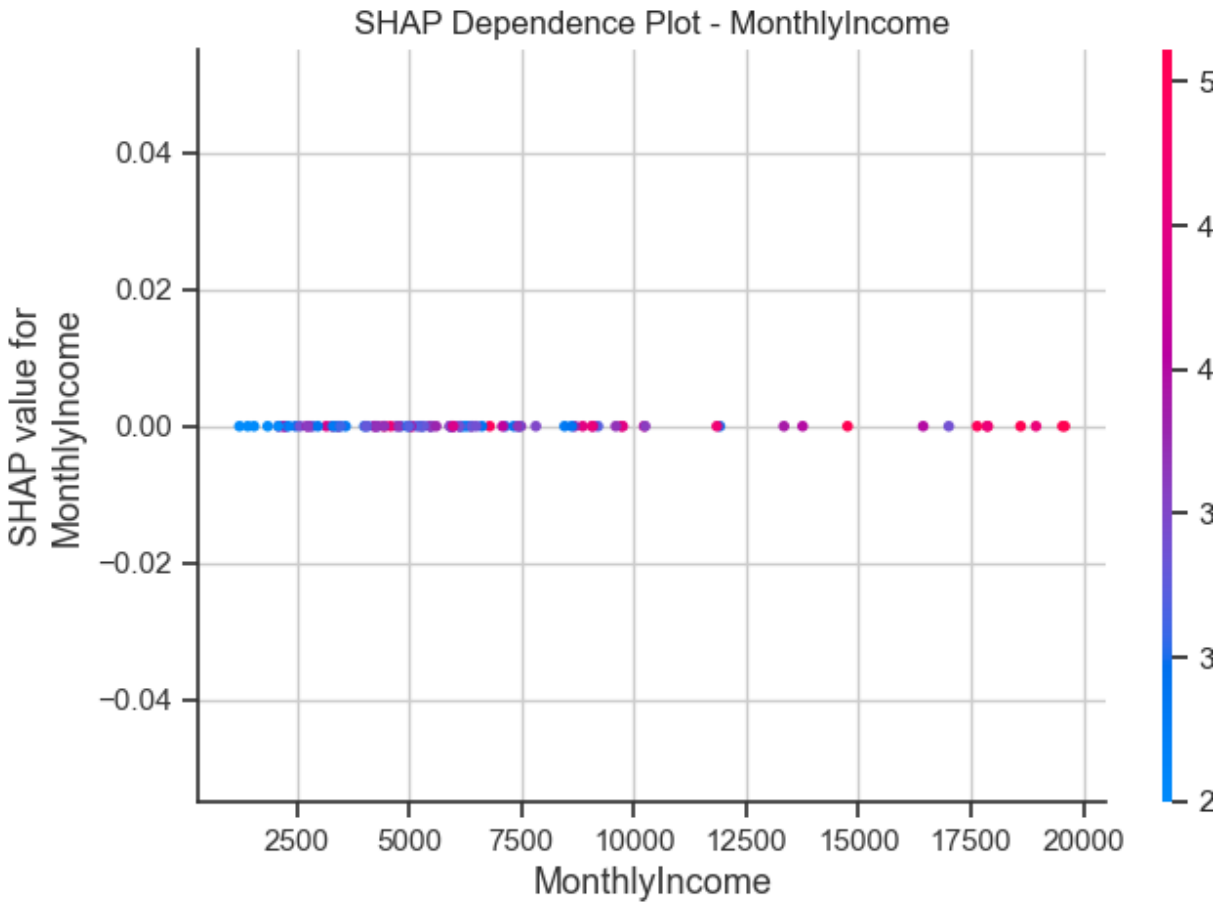




✅ `Local SHAP plots saved: shap_waterfall_plot.png & shap_local_bar_plot.png`

🔀 Dependence Plot (Feature Interaction)

Shows how attrition risk changes with one feature (e.g., MonthlyIncome) across employees, and how it interacts with another feature.



✅ Conclusions (Project 6 — SHAP Interpretability)

- **Global Drivers:** SHAP summary and feature importance plots confirm that **Distance From JobRole, and Daily Rate** are the strongest global drivers of attrition.
- **Local Explanations:** For individual employees, SHAP waterfall and bar plots show *why* a p was made, highlighting the most influential factors (e.g., frequent overtime or lower income)
- **Feature Interactions:** Dependence plots reveal how attrition risk changes across values of features (e.g., attrition risk increases at lower income bands and with overtime).

🔑 Why this matters:

- Moves the model from a **black box** → to **explainable AI**.
- Builds **trust** for HR leaders, as they can see the reasoning behind predictions.
- Supports **data-driven HR decision-making** while keeping it transparent and fair.

🚀 Next Steps

- Integrate SHAP visuals into the **Streamlit Dashboard (Project 7)** for interactive explora
- Combine SHAP outputs with **SQL pipelines (Project 8)** to deliver explainability at scale