Assignment: Feature Importance Analysis using SHAP

Domain: Education (Student Performance)

Student Name: _____

Roll Number: _____

Date: _____

Dataset: student_performance_cleaned.csv

Rows (original): 649

Rows (after de-duplication): 649

Task Type: Classification

Target Variable: pass

Preprocessing Steps

- Removed duplicates
- Missing values imputed: numeric=median, categorical=most_frequent
- Categorical encoding: One-Hot Encoding
- Train/test split: 80/20 (stratified by target)
- Feature count after encoding: 39
- Categorical columns: 0
- Numeric columns: 39

Model & Performance

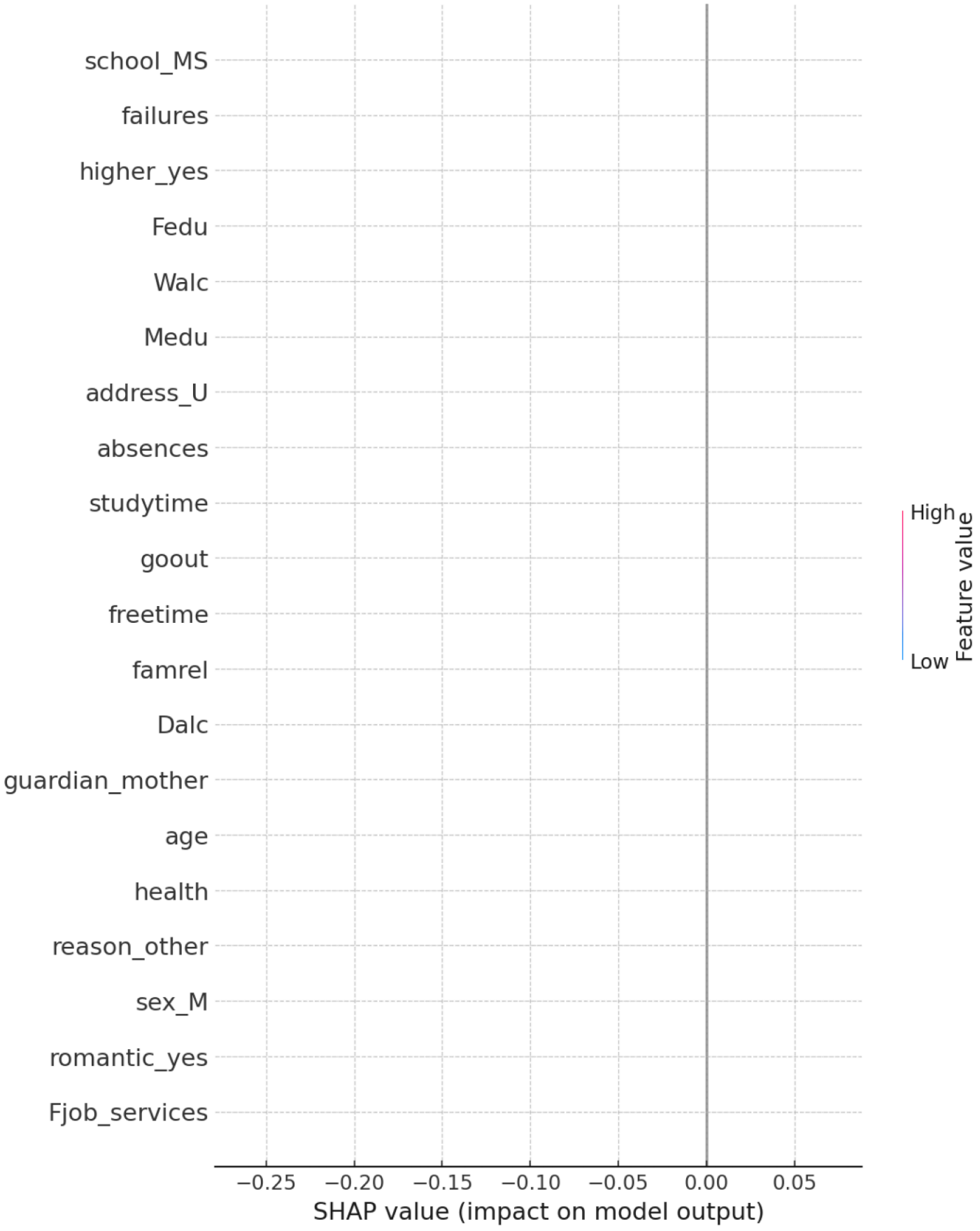
Model: RandomForestClassifier (n_estimators=300, random_state=42)

Metrics:

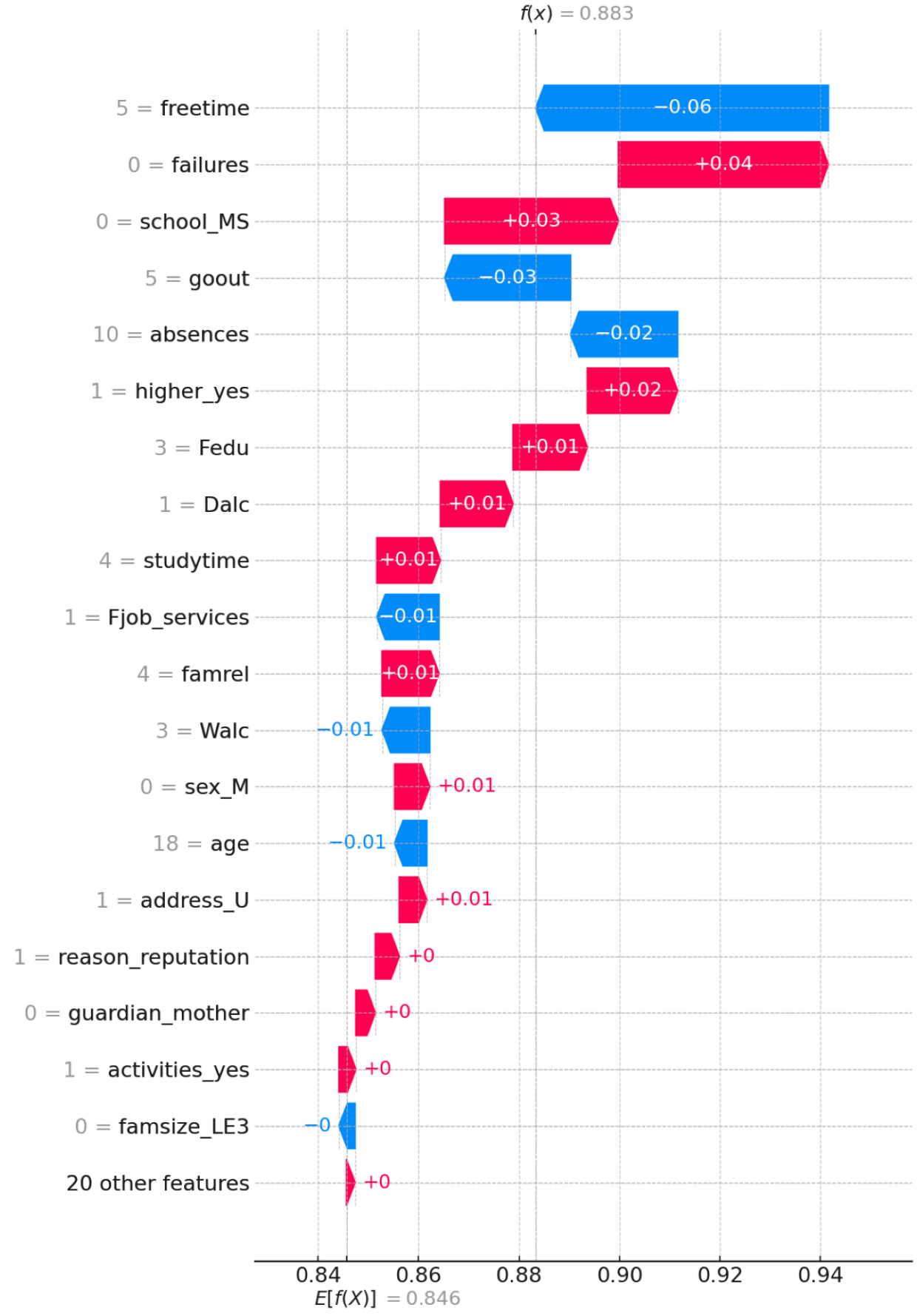
```
Task: Classification
Target: pass

Accuracy: 0.8077
Precision: 0.8571
Recall: 0.9273
F1-score: 0.8908
ROC AUC: 0.6705
```

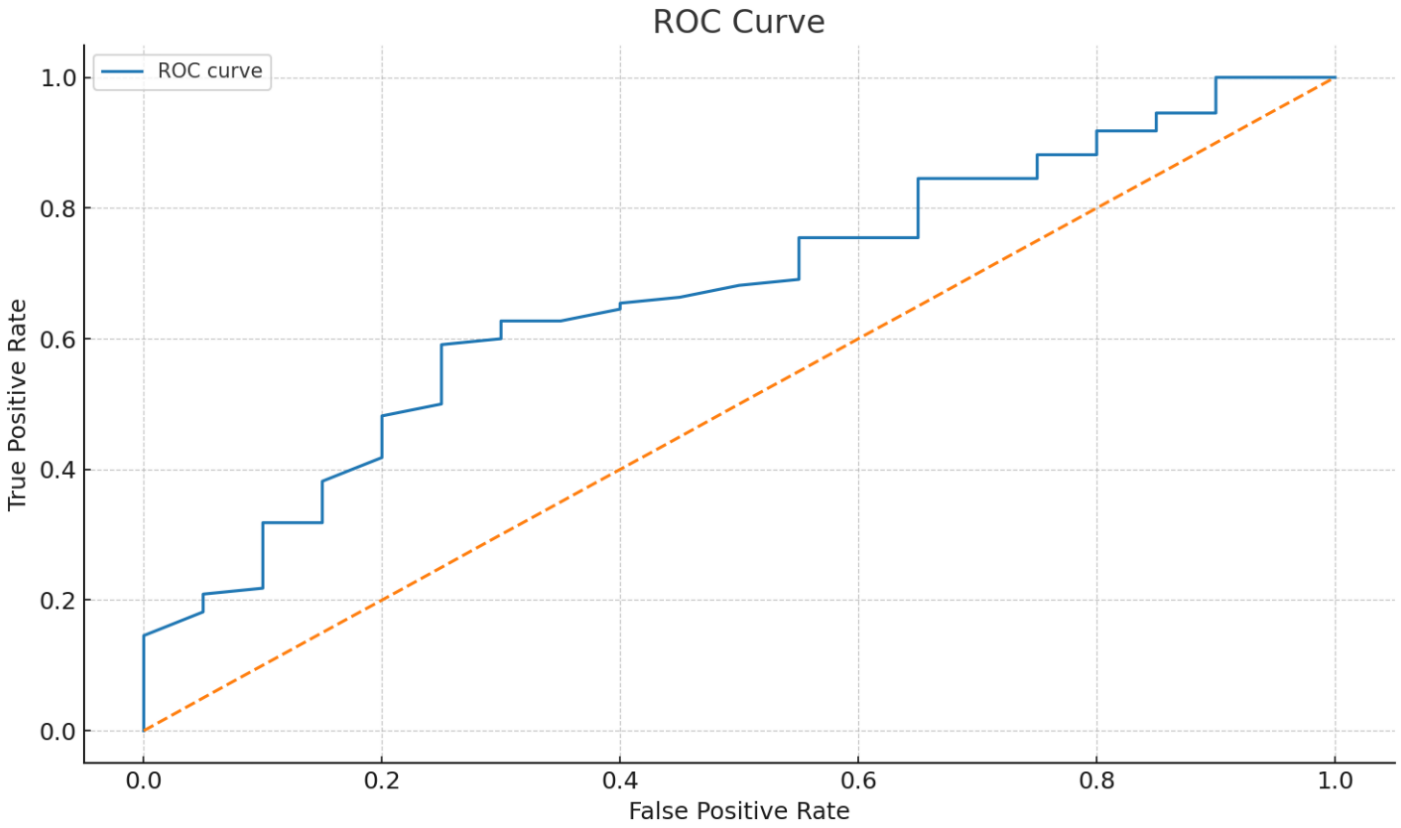
SHAP Summary Plot (overall feature importance)



SHAP Waterfall Plot (single prediction explanation)



ROC Curve


ROC Curve
ROC curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

SHAP Analysis & Interpretation

Top 5 features by SHAP importance:

1. school_MS | SHAP=0.05524 | RF=0.06437
2. failures | SHAP=0.05392 | RF=0.11896
3. higher_yes | SHAP=0.03450 | RF=0.06523
4. Fedu | SHAP=0.01734 | RF=0.04028
5. Walc | SHAP=0.01393 | RF=0.04465

Interpretation (template):

• These features show the strongest influence on predictions (by mean |SHAP|)
• Agreement with RF importance suggests consistent signal; disagreement ma
• Discuss domain sense and any surprises; note data/model limitations and ne