

Would Have Jack Survived in the Titanic



Data Bootcamp
Amina KANOUN

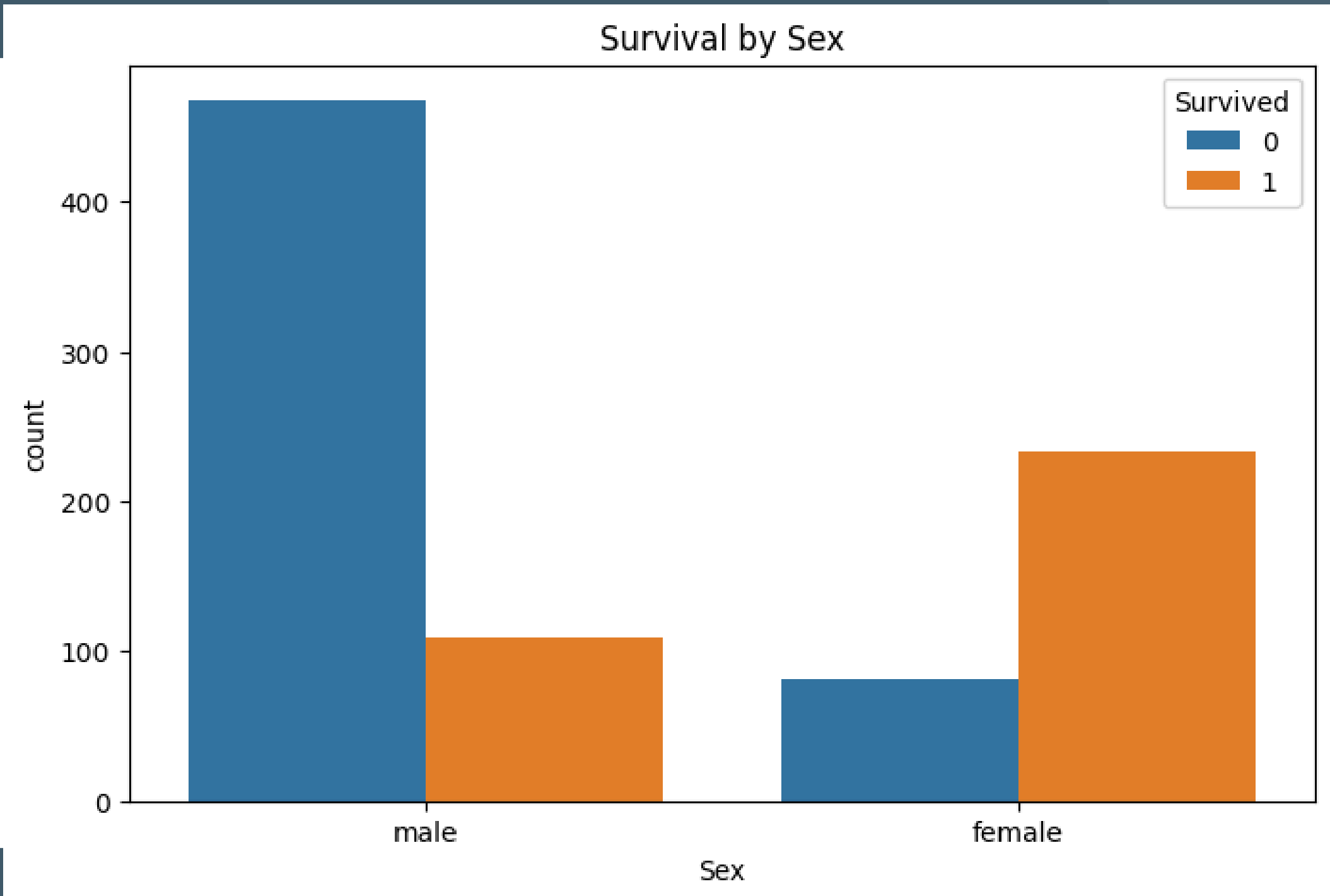
Research Question

- Initial idea: Streamlit image-generation app (changed due to technical issues)
- Inspired by a debate about the movie Titanic
- Main question:
- Could Jack have survived as a passenger?
- Focus on how class, gender, and age affect survival



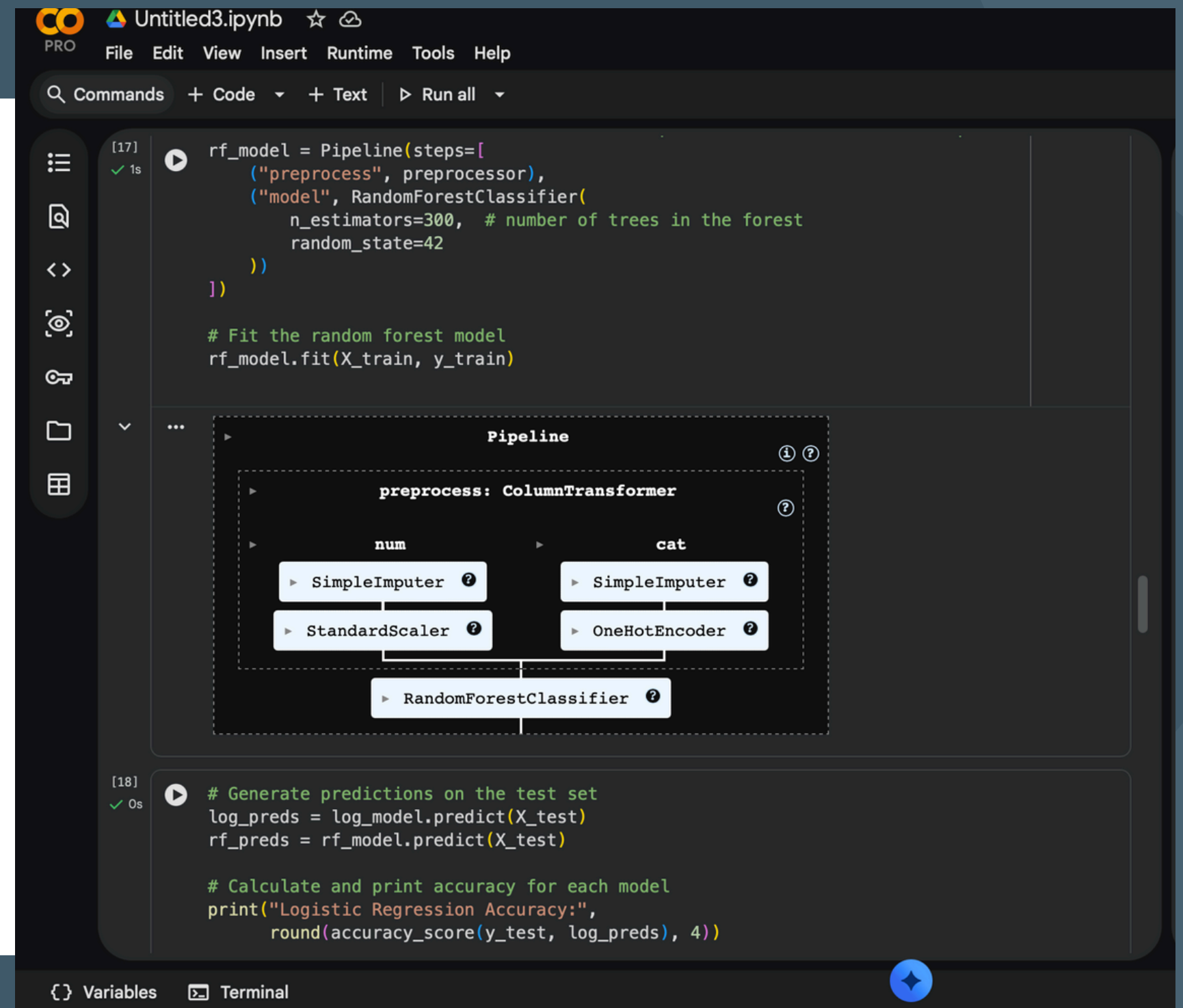
Data

- Titanic passenger dataset (Kaggle)
- 891 passengers
- Target variable:
- Survival (0 = No, 1 = Yes)
- Features used:
- Passenger class
- Sex
- Age
- Fare
- Port of embarkation



Models & Methods

- Bullets:
- Binary classification problem
- 80/20 train-test split
- Preprocessing with pipelines:
- Imputation for missing values
- One-hot encoding for categorical variables
- Models:
- Logistic Regression (baseline)
- Random Forest (ensemble model)



```
[17] ✓ 1s rf_model = Pipeline(steps=[
    ("preprocess", preprocessor),
    ("model", RandomForestClassifier(
        n_estimators=300, # number of trees in the forest
        random_state=42
    ))
])

# Fit the random forest model
rf_model.fit(X_train, y_train)
```

Diagram illustrating the Pipeline structure:

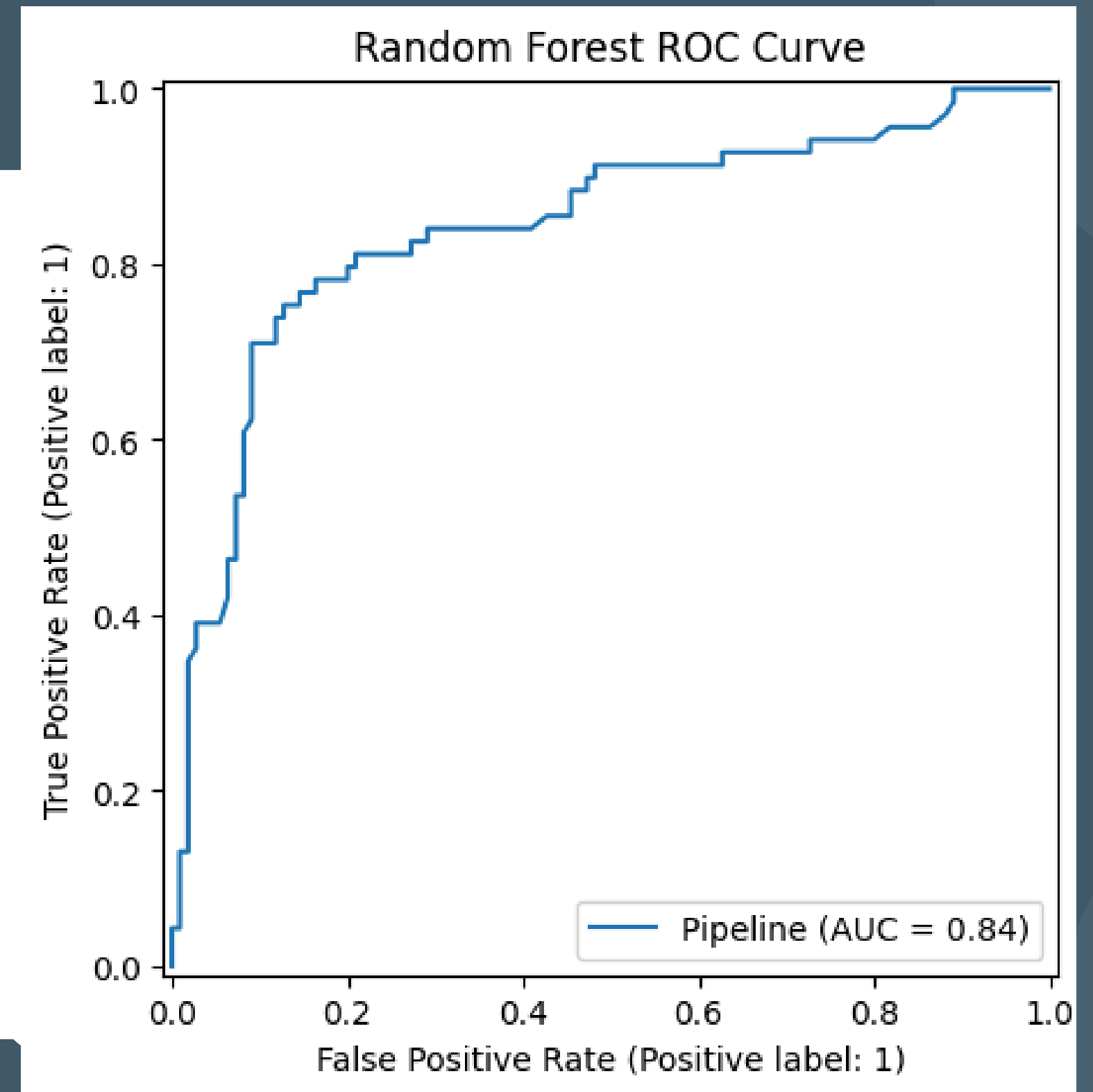
```
graph TD
    subgraph Pipeline
        subgraph preprocess [preprocess: ColumnTransformer]
            subgraph num [num]
                SI1[SimpleImputer]
                SS[StandardScaler]
            end
            subgraph cat [cat]
                SI2[SimpleImputer]
                OHE[OneHotEncoder]
            end
        end
        RF[RandomForestClassifier]
        preprocess --> RF
    end
```

```
[18] ✓ 0s # Generate predictions on the test set
log_preds = log_model.predict(X_test)
rf_preds = rf_model.predict(X_test)

# Calculate and print accuracy for each model
print("Logistic Regression Accuracy:",
      round(accuracy_score(y_test, log_preds), 4))
```

Results

- Logistic Regression accuracy \approx 0.78
- Random Forest accuracy \approx 0.82
- ROC AUC (Random Forest) \approx 0.84
- Most important factors:
 - Gender
 - Passenger class
 - Age



conclusion

- Jack-like passenger:
- Male, 20 years old
- Third class, low fare
- Predicted survival probability:
- Logistic Regression $\approx 10\%$
- Random Forest $\approx 3\%$
- “What-if” scenarios:
- If Jack were female \rightarrow survival $> 80\%$
- If Jack were 1st class \rightarrow much higher survival
- Conclusion (last bullet):
- Survival was shaped by gender and class
- Jack would have died regardless of Rose’s selfishness



