Titanic EDA + Modeling Report

Includes feature engineering and two baseline models
(Logistic Regression, Random Forest)

Source: train.csv

Dataset overview & missing values

Rows: 891 Columns: 17

Total missing values: 687

Top missing columns:

Cabin: 687

PassengerId: 0

Fare: 0

IsAlone: 0

FamilySize: 0

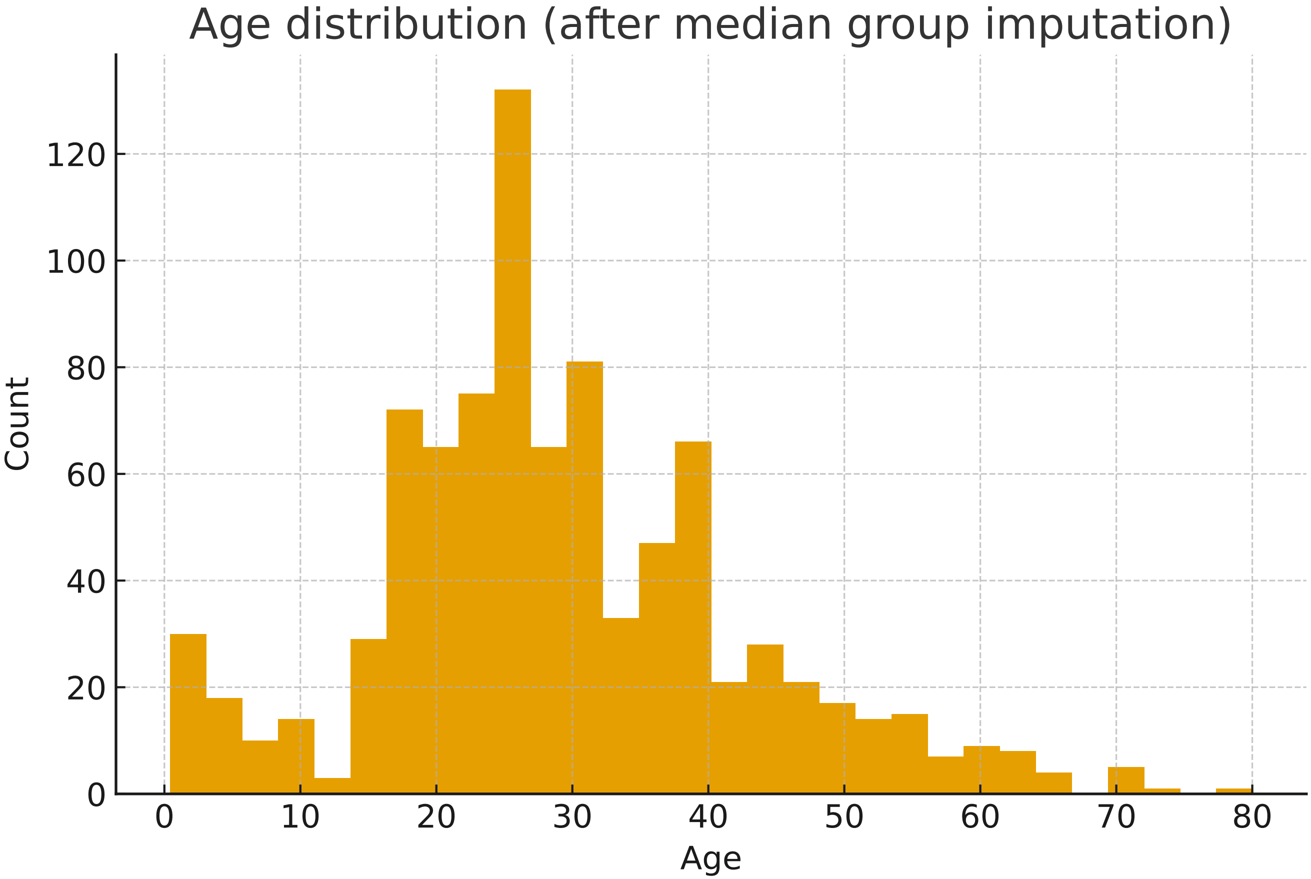
Deck: 0

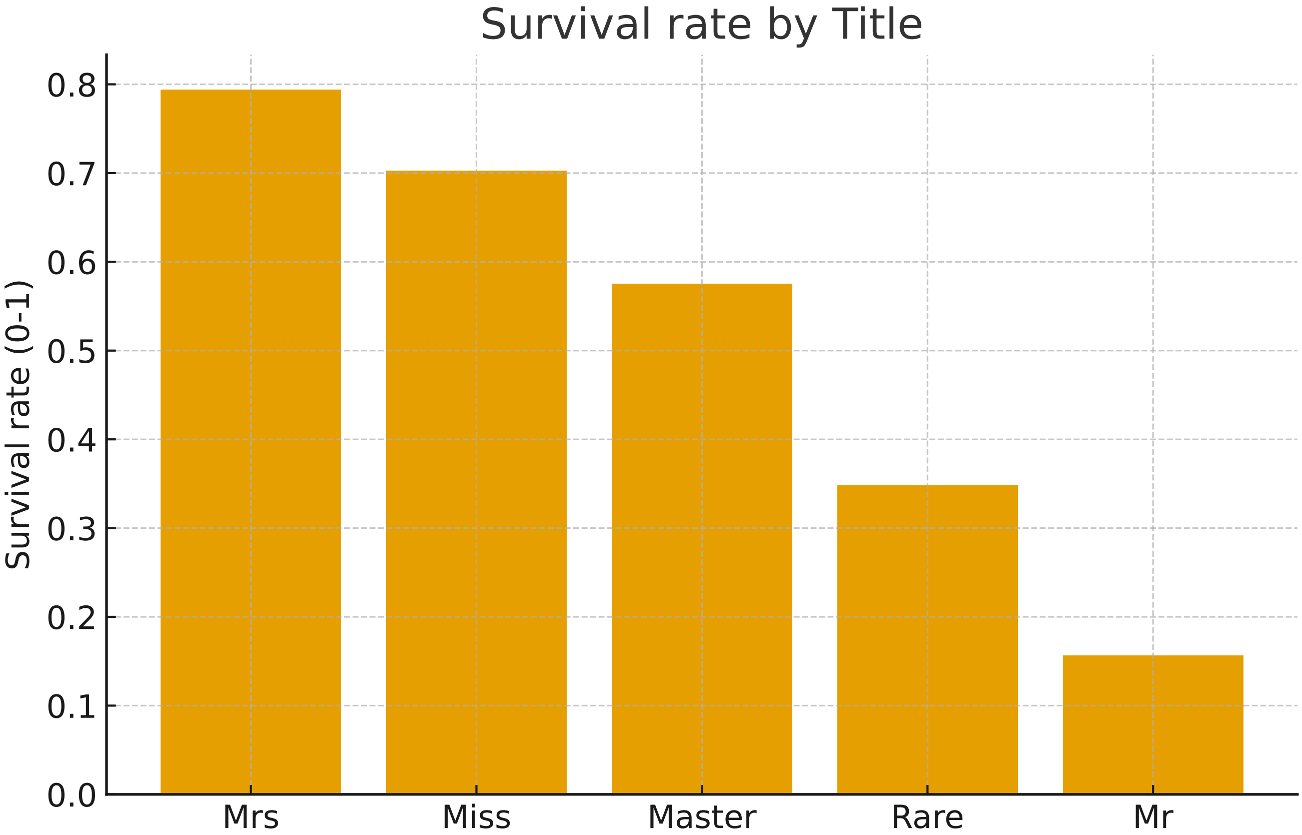
Title: 0

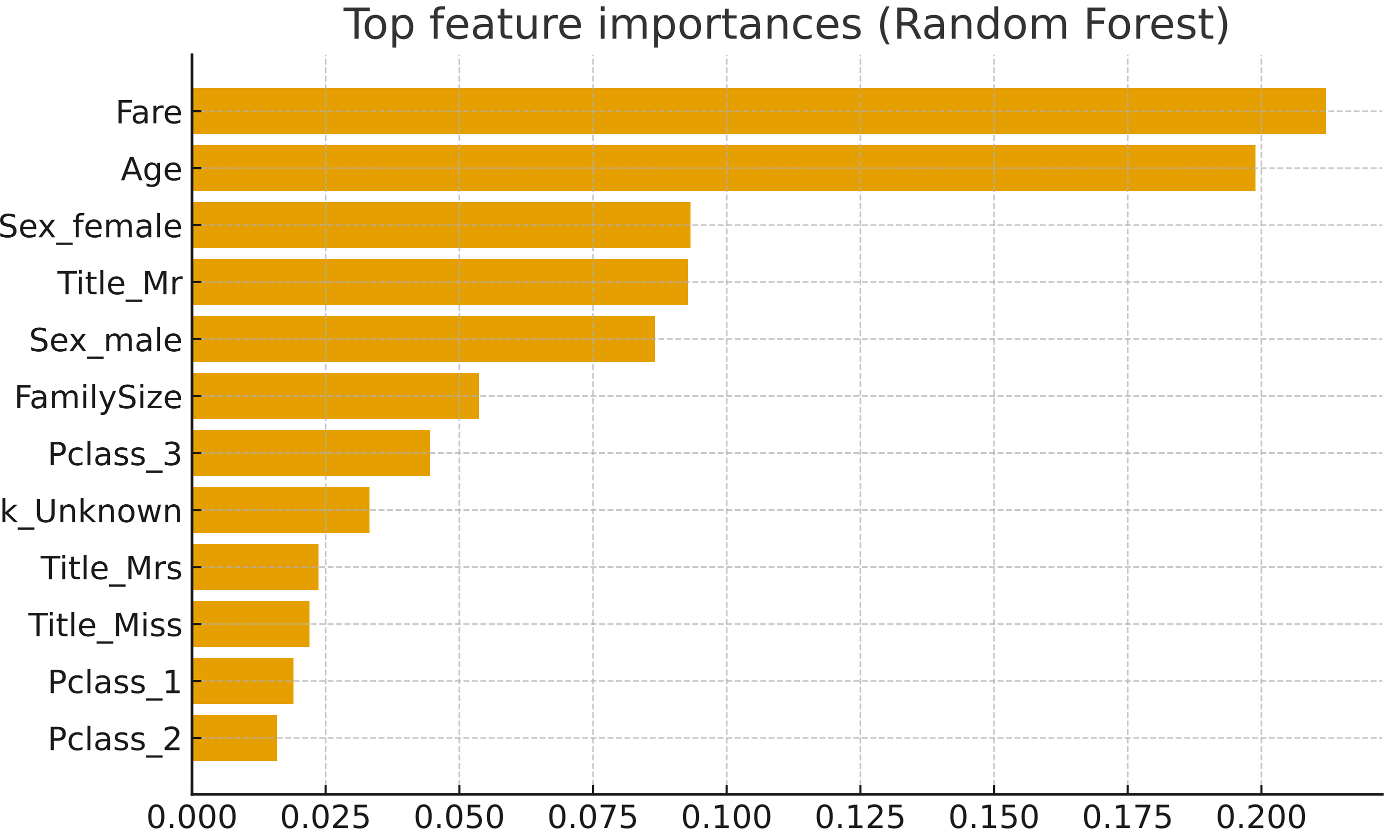
Embarked: 0

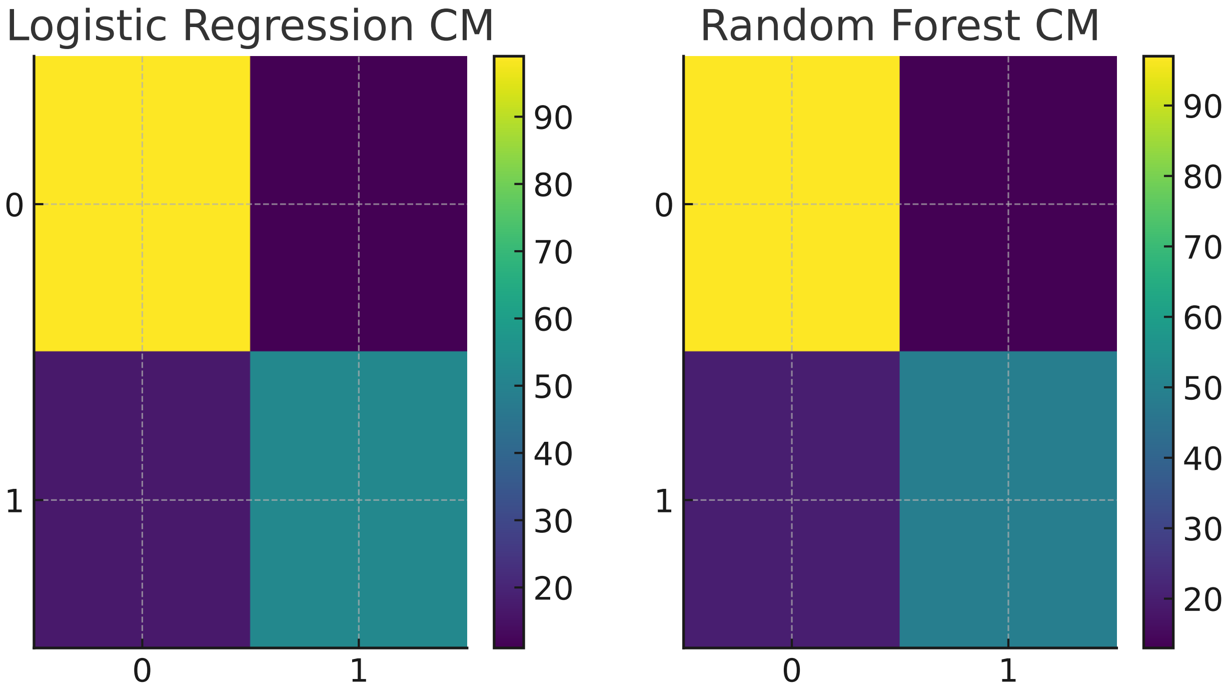
Ticket: 0

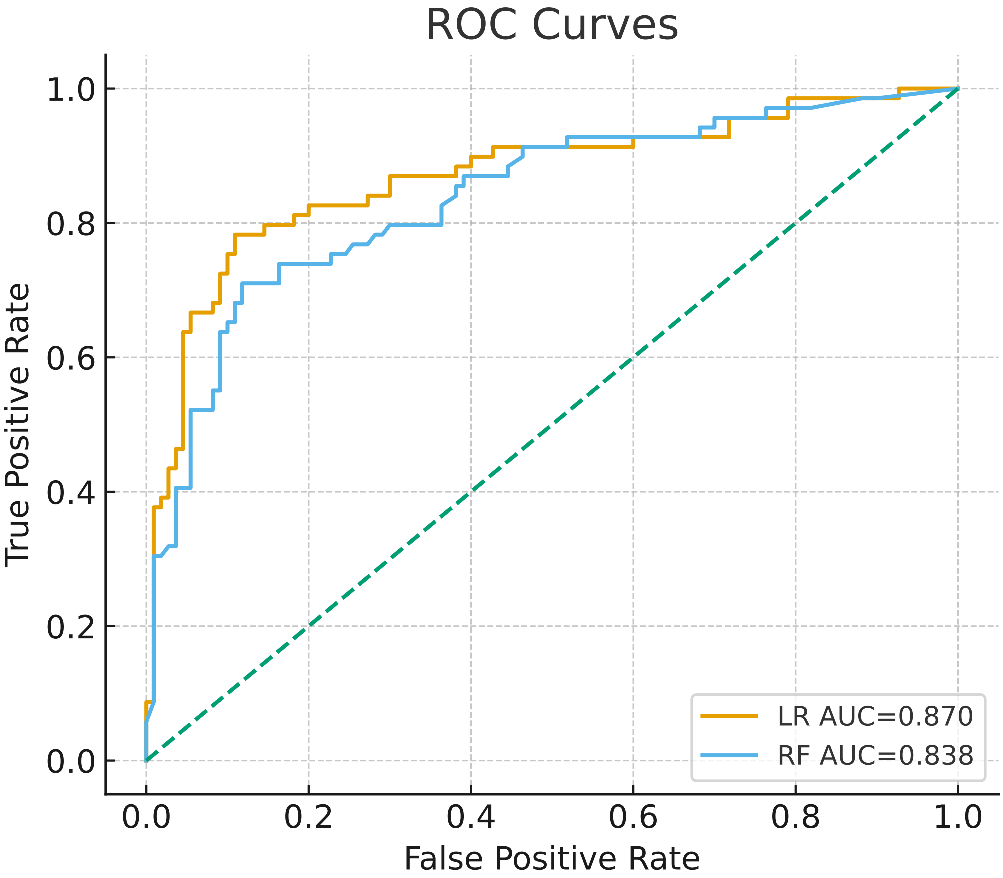
Survived: 0


Age distribution (after median group imputation)
Count
0
20
40
60
80
100
120
Age
0
10
20
30
40
50
60
70
80


Survival rate by Title
Survival rate (0-1)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Mrs
Miss
Master
Rare
Mr


Top feature importances (Random Forest)
Fare
Age
Sex_female
Title_Mr
Sex_male
FamilySize
Pclass_3
k_Unknown
Title_Mrs
Title_Miss
Pclass_1
Pclass_2
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200


Logistic Regression CM
0
1
0
1
90
80
70
60
50
40
30
20
Random Forest CM
0
1
0
1
90
80
70
60
50
40
30
20


ROC Curves
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
LR AUC=0.870
RF AUC=0.838

Observations and Recommendations:

- Logistic Regression accuracy: 0.844, ROC AUC: 0.870

- Random Forest accuracy: 0.816, ROC AUC: 0.838

- Important features (from RF): Fare, Age, Sex_female, Title_Mr, Sex_male, FamilySize

- Next steps: tune hyperparameters, cross-validate, test more feature engineering (ticket groups, inter

- For final model, consider K-fold CV and calibration of probabilities.