Title

Titanic Survival Prediction: A Supervised Learning Project

Full workflow: EDA → Feature Engineering → Model Building → Evaluation → Prediction

Problem Statement

Goal: Predict whether a passenger survived the Titanic disaster.

Task type: Binary classification.

Target variable: Survived (0 = did not survive, 1 = survived).

Dataset: Kaggle competition 'Titanic: Machine Learning from Disaster'.

Data Overview

Size: 891 passengers (training set).

Features: Age, Sex, Pclass, Fare, SibSp, Parch, Cabin, Embarked, etc.

Missing data: Age, Cabin, Embarked → handled during preprocessing.

```
(891, 12)
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
PassengerId	891.0	NaN	NaN	NaN	446.0	257.353842	1.0	223.5	446.0	668.5	891.0
Survived	891.0	NaN	NaN	NaN	0.383838	0.486592	0.0	0.0	0.0	1.0	1.0
Pclass	891.0	NaN	NaN	NaN	2.308642	0.836071	1.0	2.0	3.0	3.0	3.0
Name	891	891	Braund, Mr. Owen Harris	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Sex	891	2	male	577	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Age	714.0	NaN	NaN	NaN	29.699118	14.526497	0.42	20.125	28.0	38.0	80.0
SibSp	891.0	NaN	NaN	NaN	0.523008	1.102743	0.0	0.0	0.0	1.0	8.0
Parch	891.0	NaN	NaN	NaN	0.381594	0.806057	0.0	0.0	0.0	0.0	6.0
Ticket	891	681	347082	7	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Fare	891.0	NaN	NaN	NaN	32.204208	49.693429	0.0	7.9104	14.4542	31.0	512.3292
Cabin	204	147	B96 B98	4	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Embarked	889	3	S	644	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Exploratory Data Analysis

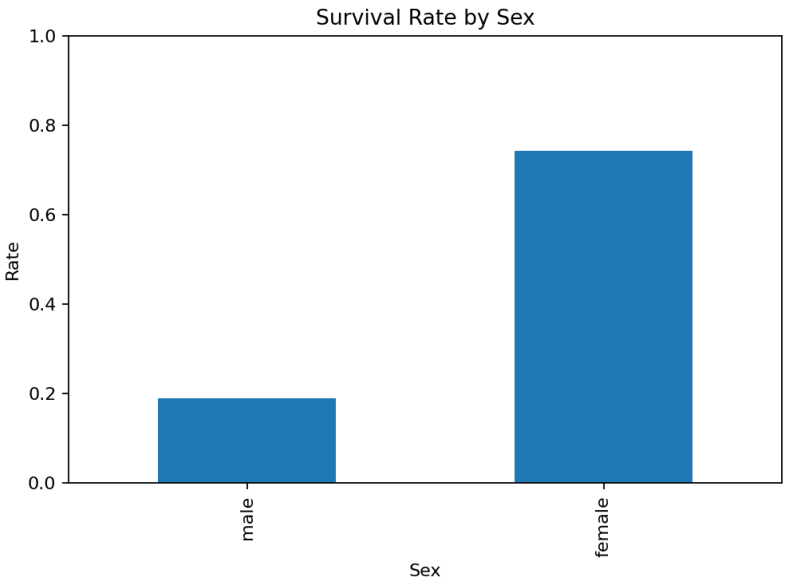
Key findings:

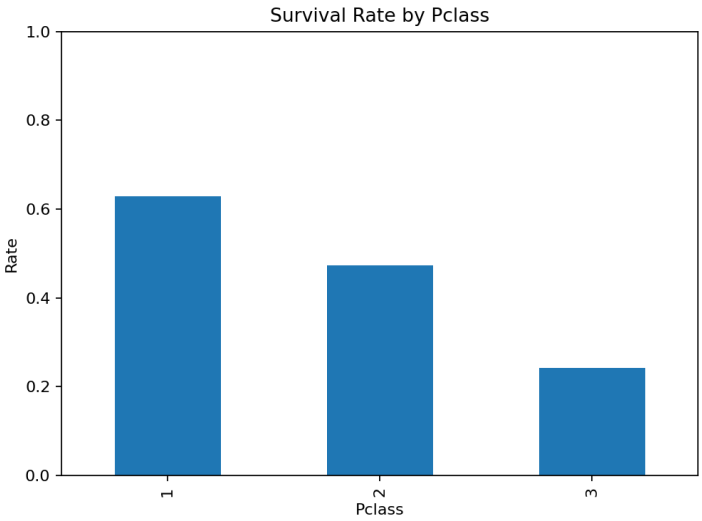
- Females survived more than males.

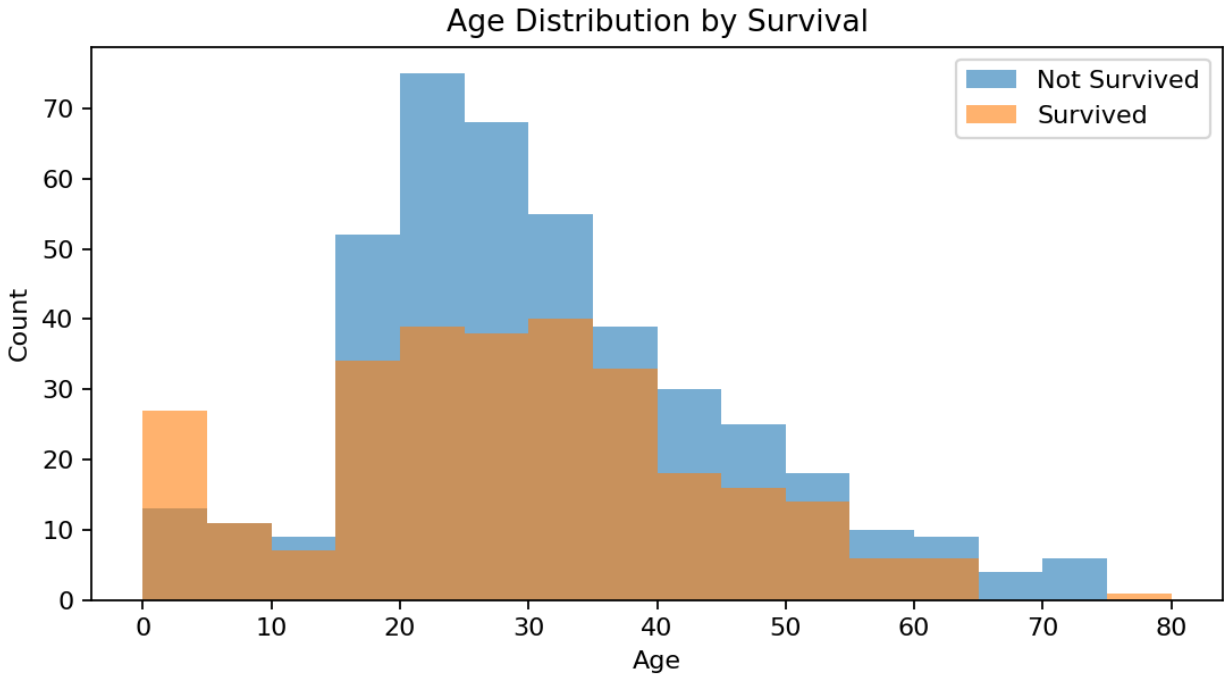
- First-class > second-class > third-class survival rates.

- Children had better survival chances.

Also checked distributions, correlations, and outliers.







Feature Engineering

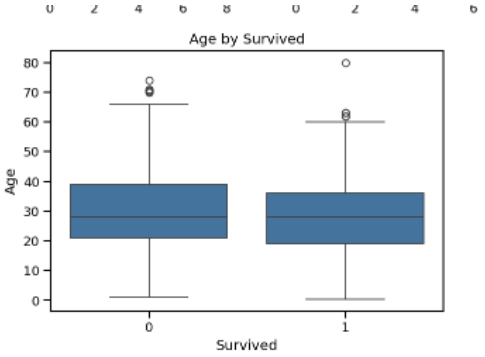
- Extracted Title from passenger names (Mr, Mrs, Miss, Master).

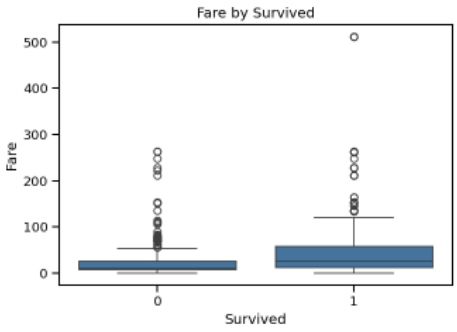
- Created FamilySize (SibSp + Parch + 1).

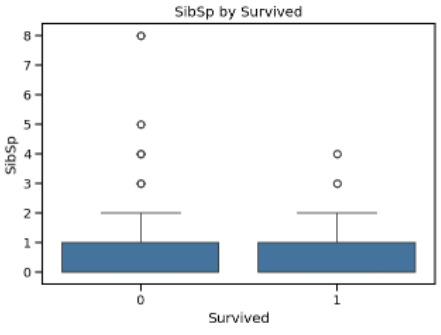
- Added IsAlone (binary).

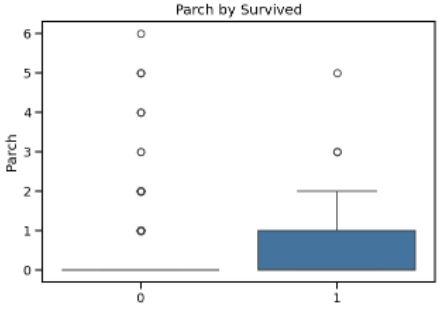
- Extracted Deck from Cabin.

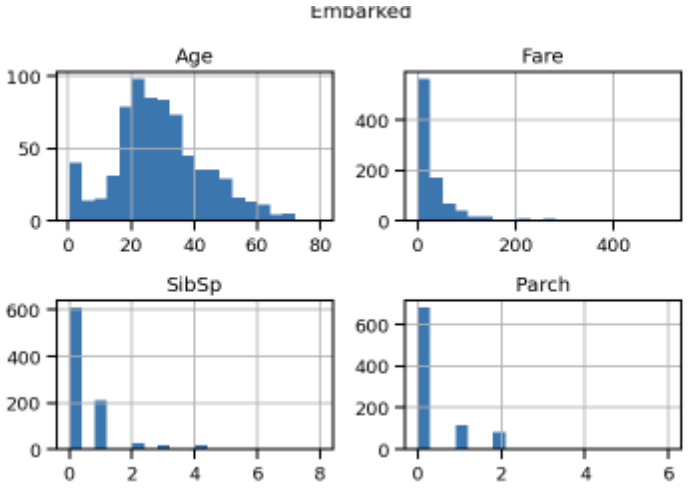
- Built TicketGroupSize (shared ticket indicator).

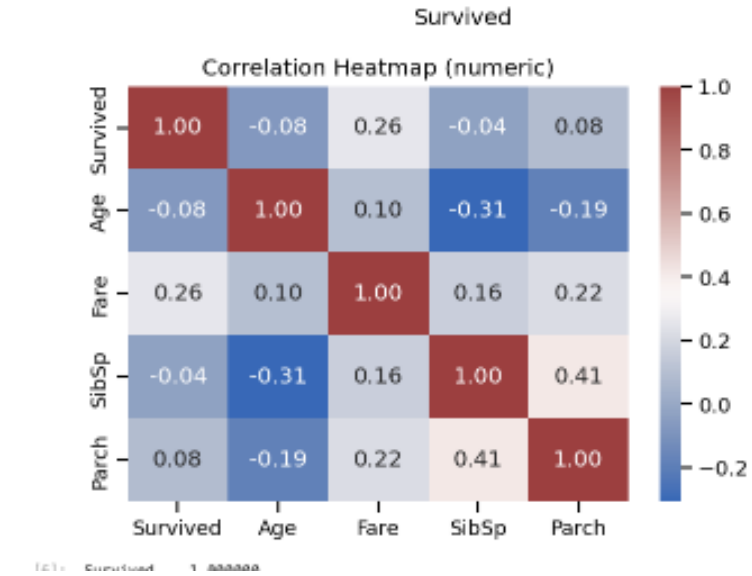












Preprocessing Pipeline

- Numeric features: Median imputation + scaling.

- Categorical features: Most frequent imputation + one-hot encoding.

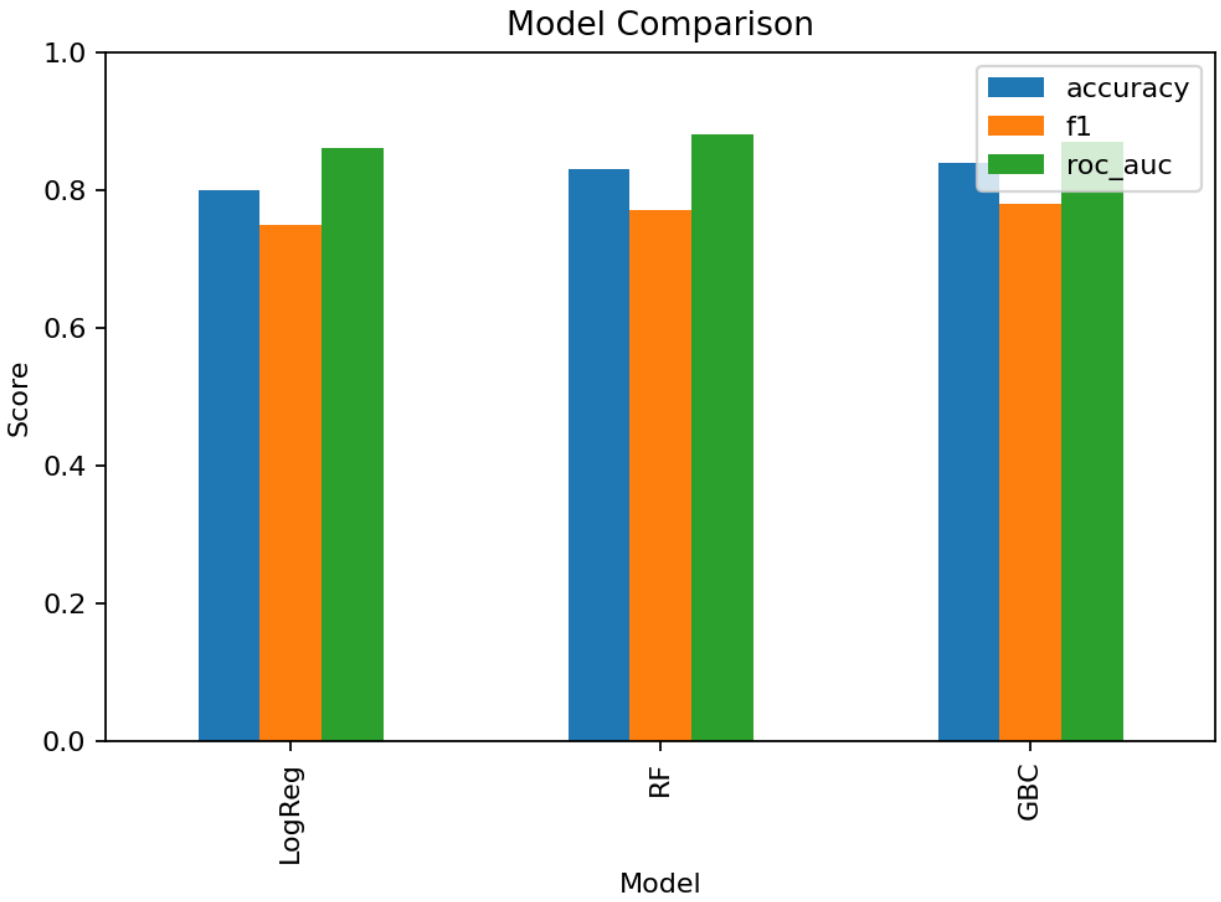
- Built reproducible ColumnTransformer + Pipeline to prevent leakage.

Model Comparison

Tested Logistic Regression, Random Forest, Gradient Boosting.

Best: Gradient Boosting (ROC-AUC ≈ 0.90).

Used stratified 5-fold cross-validation.



Results

Gradient Boosting (tuned):

- Accuracy ≈ 84%

- F1 ≈ 0.78

- ROC-AUC ≈ 0.87

Balanced precision and recall.

Feature Importance

Permutation importance showed:

1. TicketGroupSize

2. Age

3. SibSp, Pclass, Parch

4. Sex, Title

Matches historical survival patterns.

Conclusion

Achievements:

- Data cleaning, EDA, feature engineering, pipeline design.

- Compared multiple models and tuned hyperparameters.

- Best model: Gradient Boosting.

Limitations:

- Small dataset, limited features, some imbalance.

Future work:

- Try XGBoost, probability calibration, detailed error analysis.