Tianic Project

Nguyễn Toàn Thắng
Dương Tùng Thiện
Hồ Minh Tiến
Đặng Thái Tú

Nội dung trình bày

- Giới thiệu
- Công việc liên quan
- Phương pháp đề xuất
- Kinh nghiệm và kết quả
- Kết luận

1. Giới thiệu

Tổng quan dữ liệu Titanic:

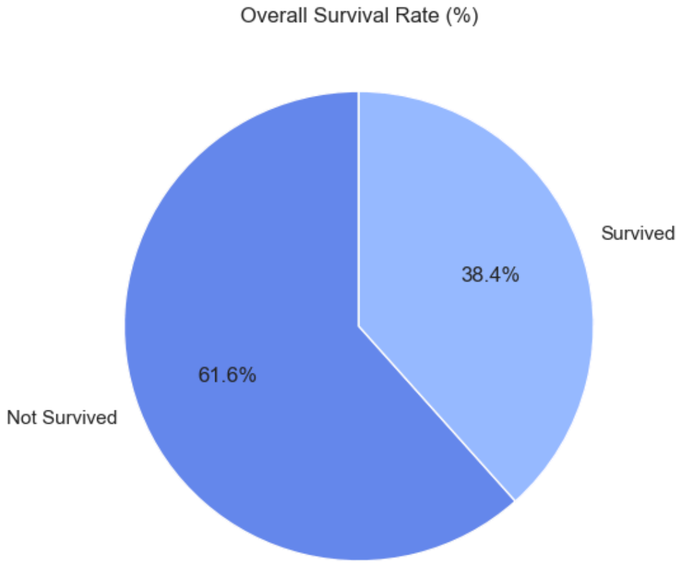
- Input: Đặc điểm cá nhân và thông tin chuyến đi của khách hàng
- Output: Khả năng sống sót (0: Chết, 1: Sống)

PassengerId	Mã hành khách	Sex	Giới tính	SibSp	Số lượng anh chị em hoặc vợ/chồng đi cùng
Survived	Kết quả sống sót (0 = ❌, 1 = ✅)	Age	Tuổi của hành khách	Parch	Số lượng cha/mẹ hoặc con đi cùng
Pclass	Hạng vé của hành khách (1, 2, 3)	Ticket	Mã vé	Cabin	Số hiệu phòng trên tàu
Name	Họ tên hành khách	Fare	Giá vé	Embarked	Cảng lên tàu (C = Cherbourg, Q = Queenstown, S = Southampton)

1. Giới thiệu

Thách thức:

- Dữ liệu chứa nhiều giá trị bị thiếu (missing value)
- Dữ liệu bị mất cân bằng giữa số người sống và tử vong.
- Nhiều biến định tuyến cần được mã hóa (Sex, Pclass, Embarked,...)
- Mô hình dữ liệu nhỏ, dễ dẫn đến overfitting




```
===== Number of missing values (train dataset) =====

Age          177
Cabin        687
Embarked       2


===== Number of missing values (test dataset) =====

Age           86
Cabin        327
Fare           1
```

2. Công việc liên quan

Classical Models:

- Logistic Regression – mô hình tuyến tính cơ bản cho bài toán nhị phân

Emsemble Methods:

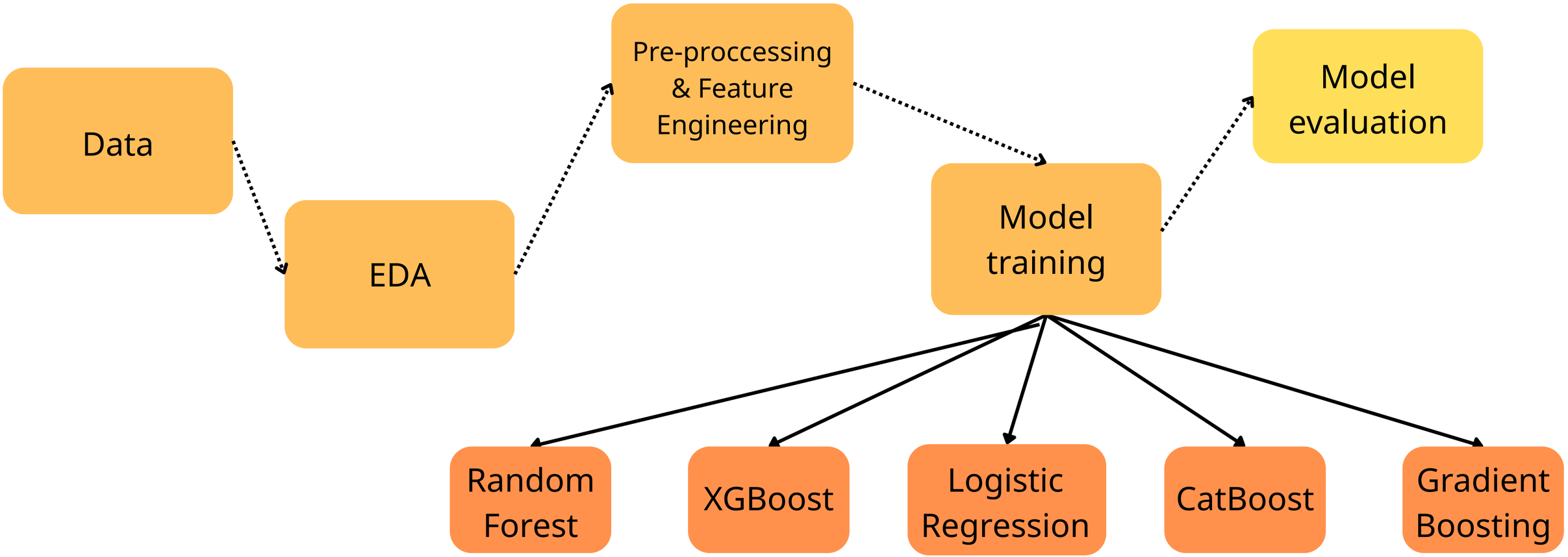
- Random Forest, Gradient Boosting (XGBoost) và Gradient Boosting, CatBoost– tăng độ chính xác bằng cách kết hợp nhiều cây quyết định.
- Các mô hình này đã được chứng minh là hiệu quả trên Titanic Dataset của Kaggle.

Feature Engineering Approaches:

- Tạo đặc trưng mới từ các cột như FamilySize, AgeAndPclass, FareTransformed.
- Xử lý dữ liệu bị thiếu bằng median, mode,...

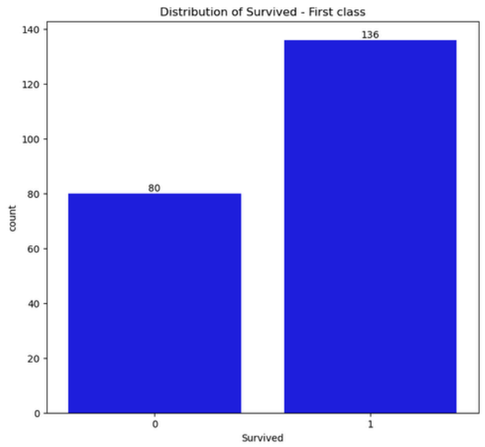
3. Phương pháp đề xuất

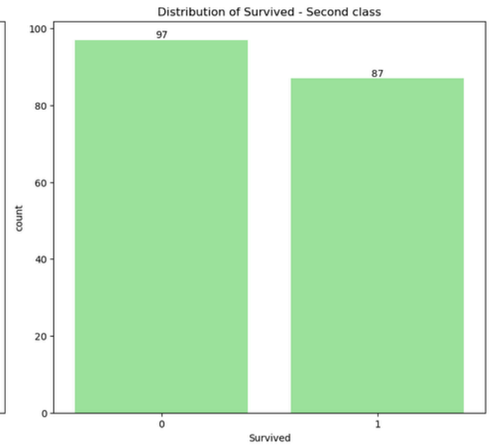
Quá trình chính:

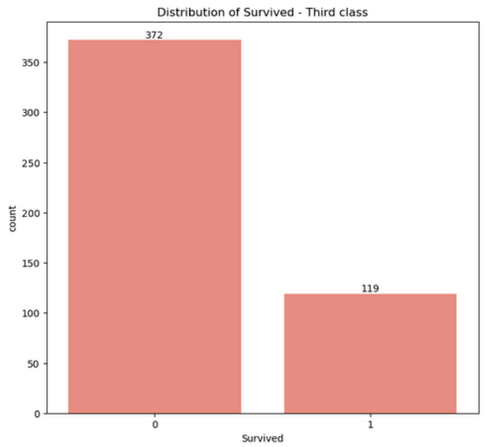


3. Phương pháp đề xuất

Exploratory data analysis (EDA)





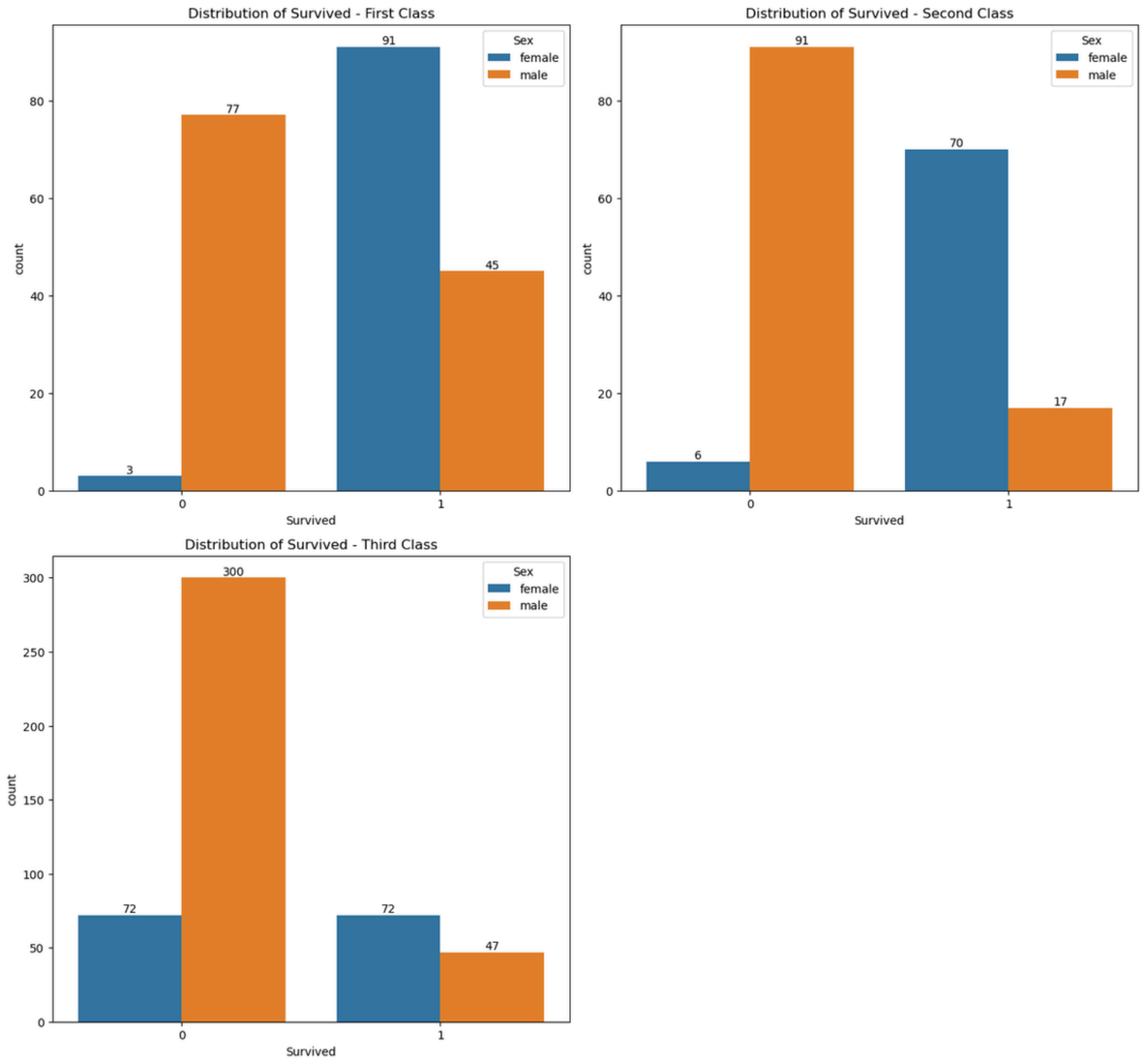


- First Class: Có tỷ lệ sống cao nhất trong ba hạng vé.
- Third Class: Số người tử vong vượt xa số người sống sót.
- Hạng vé là yếu tố quan trọng ảnh hưởng đến khả năng sống sót của hành khách.

3. Phương pháp đề xuất

Exploratory data analysis (EDA)

- Tỷ lệ sống sót của nữ cao hơn nam do quy tắc “phụ nữ và trẻ em được ưu tiên”
- Là yếu tố quan trọng góp phần tăng khả năng dự đoán khả năng sống sót



3. Phương pháp đề xuất

Tiền xử lý dữ liệu

Xử lí các giá trị bị thiếu:

- Cabin: Loại bỏ đặc trưng.
- Age: điền bằng median của toàn bộ tập dữ liệu, trừ nhóm “Master” — dùng median riêng vì nhóm này chỉ gồm người dưới 12 tuổi.
- Fare : Điền bằng median của nhóm Pclass tương ứng.
- Embarked : Điền bằng giá trị “S” vì là điểm lên tàu chính là Southampton.

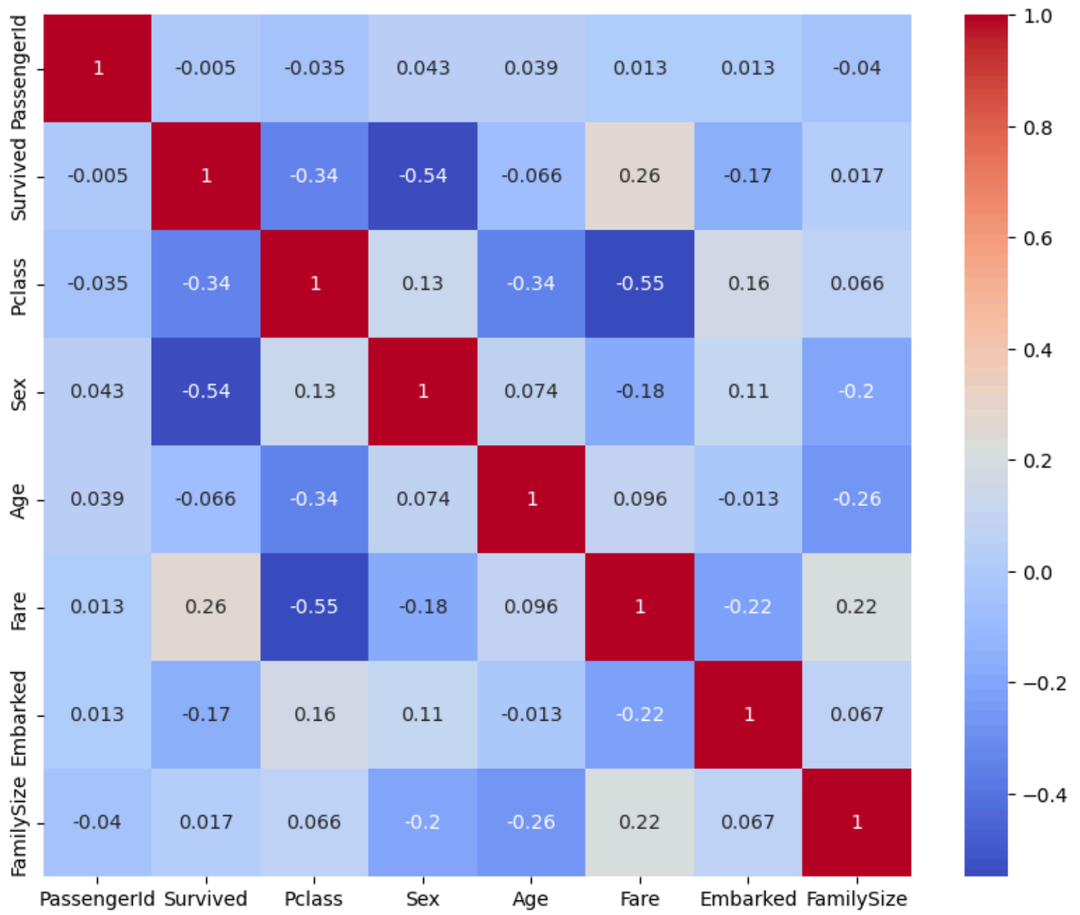
	Missing Value	**Count**
Test Dataset	Age: 20.57% Cabin: 78.23% Fare: 0.24%	177 687 2
Train Dataset	Age: 19.87% Cabin: 77.1% Embarked: 0.22%	86 327 1

3. Phương pháp đề xuất

Feature engineering

- FamilySize = SibSp + Parch + 1 (bao gồm cả hành khách) => Giúp tinh gọn lại tập dữ liệu, cải thiện hiệu suất huấn luyện mô hình.

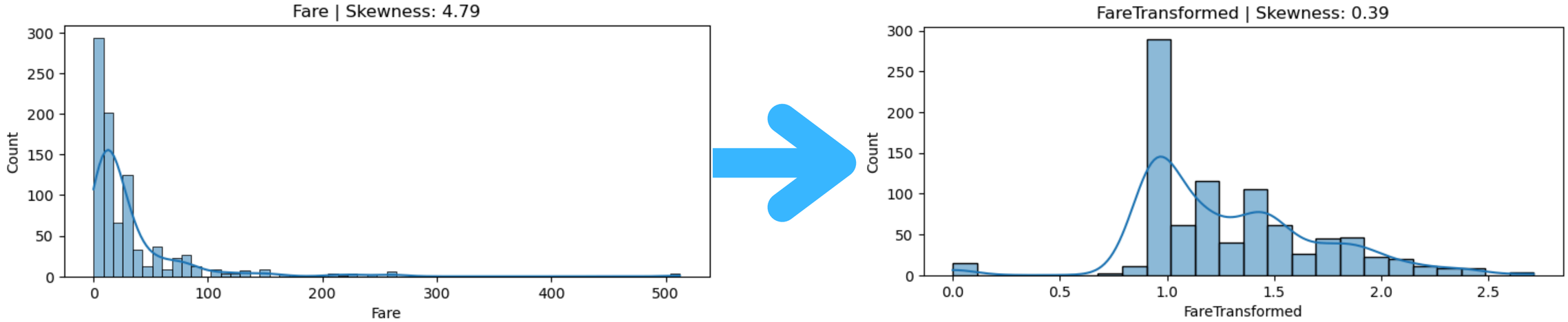
- AgeAndPclass = Age × Pclass: Giá trị càng thấp, tỉ lệ sống sót càng cao.



3. Phương pháp đề xuất

Feature engineering

- Sự phân bổ của tính năng Fare bị lệch nhiều (nghiêng phải).
- Quyết định thực hiện chuyển đổi cho tính năng Fare
- Tạo tính năng mới có tên FareTransformed

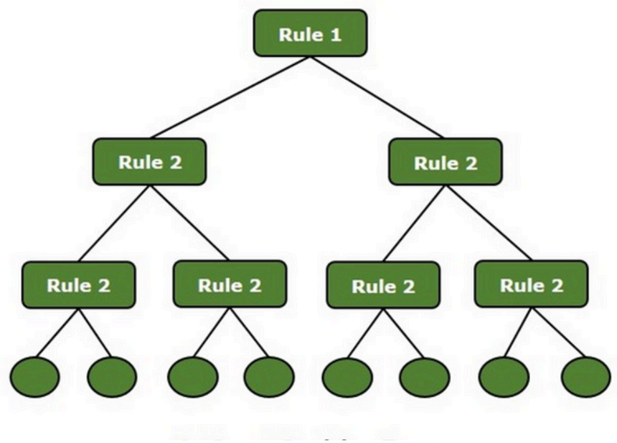


3. Phương pháp đề xuất

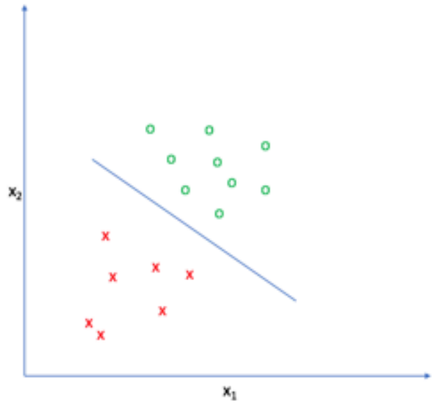
Model training:

Các mô hình thử nghiệm gồm:

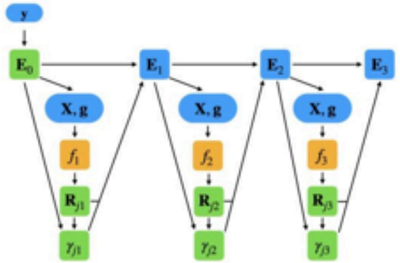
- Cơ bản: Logistic Regression, Gradient Boosting, CatBoost
- Ensemble: Random Forest, XGBoost



CatBoost

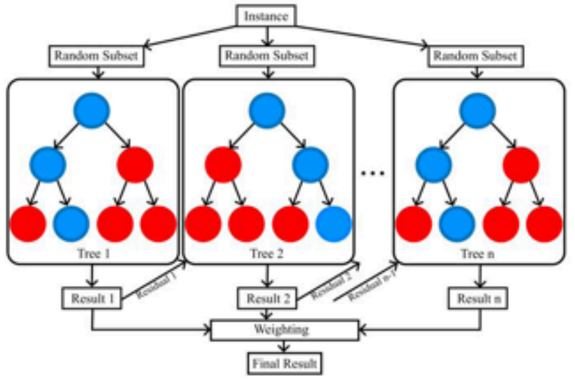


Logistic Regression



Gradient Boosting

Random Forest



XGBoost

3. Phương pháp đề xuất

Đánh giá mô hình

Đánh giá bằng Cross-Validation:

- Áp dụng 5-fold Cross-Validation để đánh giá độ ổn định và khả năng tổng quát của mô hình.
- Giữ tỷ lệ sống sót (38%) cân bằng trong mỗi tập để đảm bảo đánh giá công bằng và chính xác hơn.
- Kết quả cuối cùng được lấy trung bình từ 5 lần huấn luyện và kiểm thử.

4. Kinh nghiệm và kết quả

Kết quả của các lần thực nghiệm:

Feature Engineering	Best Model	Accuracy	F1	Precision	Recall	Kaggle
Baseline	XGBoost & CatBoost	0.8204	0.7515	0.7945	0.7148	0.76794
FamilySize	Random Forest	0.8339	0.7651	0.8291	0.7117	0.77272
AgeAndPclass	CatBoost	0.8204	0.7510	0.7921	0.7173	0.77033
ScaledData	**XGBoost**	**0.8215**	**0.7578**	**0.7868**	**0.7325**	**0.78229**

Mô hình đạt được hiệu suất cao nhất trên Kaggle: XGBoost với tỉ lệ 0.78229%

4. Kinh nghiệm và kết quả

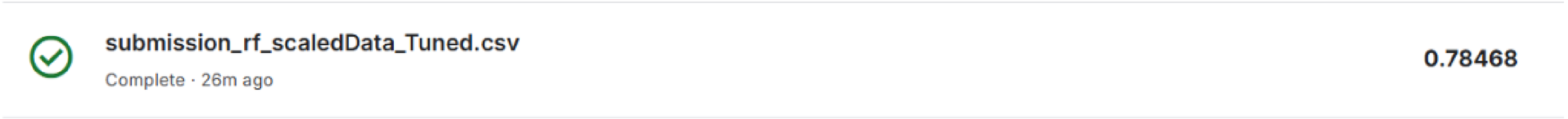
Kết quả của scaledData sau khi cải tiến:

Model	Accuracy	F1	Precision	Recall	Kaggle
Random Forest	**0.8226**	**0.7568**	**0.7910**	**0.7260**	**0.78468**
Logistic Regression	0.7980	0.7051	0.7925	0.6359	0.77990
XGBoost	0.8227	0.7576	0.7884	0.7323	0.77751
Gradient Boosting	0.8215	0.7542	0.7937	0.7937	0.77751
CatBoost	0.8215	0.7542	0.7932	0.7199	0.76555

Kết quả ổn định (Accuracy khoảng 0.79–0.82), trong đó Random Forest đạt hiệu suất cao nhất

5. Phần kết luận

- Bất chấp những thách thức đáng kể đặt ra trong suốt nghiên cứu này, các mô hình được phát triển đã đạt được kết quả.
- Mặc dù điểm số này chưa đáp ứng được mục tiêu nghiên cứu chính nhưng nó vẫn thể hiện sự tiến bộ đáng kể của dự án.



Thank You