Analysis of the Titanic Disaster Using Machine Learning Models

PHÂN TÍCH THẢM HỌA TÀU TITANIC SỬ DỤNG MÔ HÌNH MÁY HỌC

Lê Thanh Phát, Nguyễn Hữu Tri, Lư Hồng Phúc, Đỗ Duy Quý

GIỚI THIỆU

Bối cảnh

Thảm họa Titanic là một case study kinh điển để ứng dụng Khoa học dữ liệu, nhằm tìm ra các yếu tố quyết định sự sống còn từ dữ liệu lịch sử

Mục tiêu chính:

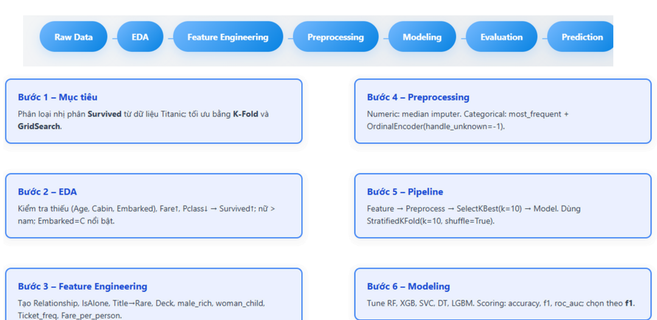
- Xây dựng mô hình dự đoán khả năng sống sót **(Survived: 1 = Sống, 0 = Tử vong)** với độ chính xác cao nhất có thể.
- Xác định các yếu tố then chốt ảnh hưởng đến sự sống còn, như Age, Sex, và Pclass.
- Đánh giá và so sánh hiệu suất của các thuật toán để tìm ra phương pháp tối ưu cho bộ dữ liệu này

DỮ LIỆU

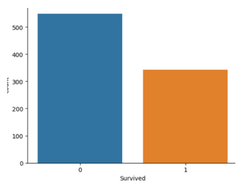
- "Titanic: Machine Learning from Disaster" từ Kaggle, gồm 891 mẫu huấn luyện và 418 mẫu kiểm thử

Nguồn dữ liệu: Kaggle: https://www.kaggle.com/competitions/titanic/data

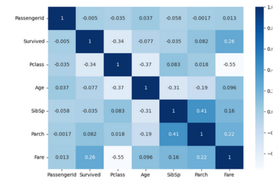
PHƯƠNG PHÁP ĐỀ XUẤT



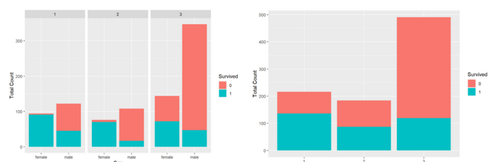
1.Phân tích & Khám phá dữ liệu (EDA)



Đầu ra chênh lệch rõ rệt

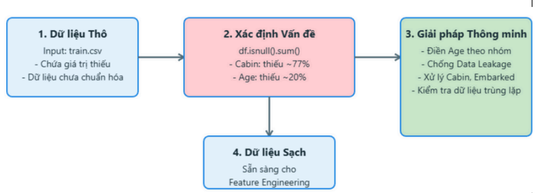


Tương quan của toàn bộ các đặc trưng

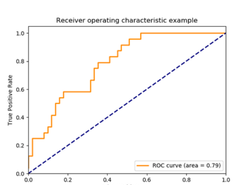


Tỉ lệ sống sót phụ thuộc nhiều vào giới tính và hạng vé

2. Quy trình tiền xử lý dữ liệu



4.Đánh giá & Valiadation



Kết hợp cả 3 phương pháp đánh giá:

- ***Accuracy***
- ***F1***
- ***ROC - AUC***

ACTUAL VALUES	POSITIVE	NEGATIVE
POSITIVE	TP	FN
NEGATIVE	FP	TN

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

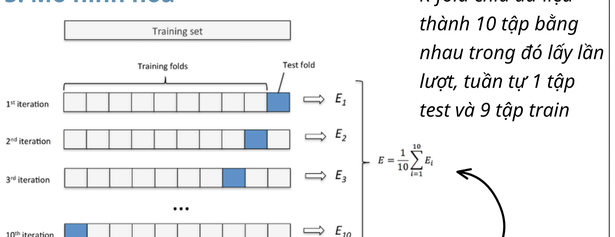
3. Kỹ thuật đặc trưng

Trích xuất thông tin từ dữ liệu văn bản

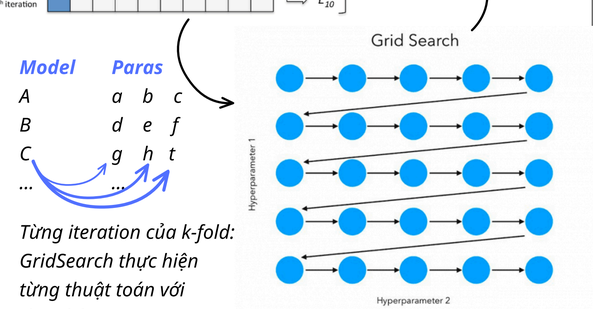
Tổng hợp và tạo đặc trưng dựa trên giả thuyết

Từ những phỏng đoán và phân tích để tạo ra các biến số có sức mạnh dự báo.

5. Mô hình hóa



K-fold chia dữ liệu thành 10 tập bằng nhau trong đó lấy lần lượt, tuần tự 1 tập test và 9 tập train



Từng iteration của k-fold: GridSearch thực hiện từng thuật toán với từng bộ parameter tương ứng

THỬ NGHIỆM VÀ KẾT QUẢ

Thử nghiệm	Cải tiến chính	Mô hình tốt nhất	Score (Kaggle)
Lần 1	Baseline: Xử lý dữ liệu cơ bản, xóa Cabin & Name.	Random Forest	0,7559
Lần 2	Giữ Cabin, tạo đặc trưng Relationship.	Random Forest	0,7775
Lần 3	Trích xuất Title & Deck.	XGBoost	0,75837
Lần 4	Thử nghiệm VotingClassifier với baseline.	VotingClassifier	0,77272
Lần 5	VotingClassifier (4 mô hình) + Features nâng cao.	VotingClassifier	0,78229

Đánh giá:

- Mô hình đã cho ra dự đoán đúng khoảng 80%
- Mới chỉ khai thác và phát hiện được một phần thông in bên trong dữ liệu.
- Chưa hiểu rõ được lịch sử tại thời gian lúc tàu Titanic gặp tai nạn và chưa hiểu rõ được văn hóa lúc đó để đánh giá và liên kết dữ liệu tốt hơn.
- Các mô hình và phương pháp làm còn đơn sơ và chưa đi sâu vào tinh chỉnh mô hình phù hợp.

HƯỚNG CẢI TIẾN MÔ HÌNH

Hướng 1: Hiểu rõ dữ liệu, lịch sử và văn hóa tại thời gian lúc tàu Titanic gặp tai nạn và cấu tạo của tàu, nhằm:

- Xử lí dữ liệu phù hợp hơn
- Cách điền giá trị thiếu thông minh hơn
- Chuẩn hóa, rời rạc hóa dữ liệu
- Thiết kế đặc trưng tốt hơn
- Dùng PCA, T-sne,... để trích xuất
- Dùng Chi2, ANOVA,.. để lựa chọn những đặc trưng phù hợp

Hướng 2: Phân tích sâu hơn và đánh giá chính xác hơn từng mô hình, nhằm:

- Lựa chọn được mô hình phù hợp với dữ liệu và phương pháp đang sử dụng:
- Logistic Regression, Genetic Algorithms
- Tinh chỉnh sâu vào mô hình
- Thử nghiệm các mô hình mới.

Hướng 3: Sử dụng phương pháp mới.