Challenge 1: Titanic – Machine Learning from Disaster

Trần Quốc Hoàng, Trần Chí Vỹ,

Phan Trần Hữu Tấn

Trường đại học Sài Gòn, Việt Nam

Tác giả liên hệ(s) E-mail(s): 3123580065@sv.sgu.edu,vn

Tác giả đóng góp: 3123580015@sv.sgu.edu.vn;

3123580042@sv.sgu.edu.vn;

Các tác giả này đóng góp như nhau trong cho công trình này.

Tóm tắt

Thảm họa Titanic vẫn là một nghiên cứu điển hình quan trọng trong lĩnh vực quản lý an toàn hàng hải, khiến việc dự đoán khả năng sống sót của hành khách trở thành nhiệm vụ then chốt để kiểm tra các yếu tố rủi ro. Nghiên cứu này tận dụng các phương pháp Học máy (Machine Learning) và Khai phá dữ liệu (Data Mining) để dự báo kết quả sống sót và xác định các mô hình liên kết trạng thái sống sót với các yếu tố quyết định chính. Sử dụng bộ dữ liệu Kaggle bao gồm 12 thuộc tính và 342 người sống sót, nghiên cứu đã khám phá các quy trình khai phá dữ liệu, bao gồm phân tích dữ liệu khám phá (Exploratory Data Analysis-EDA), huấn luyện mô hình và đánh giá nghiêm ngặt. Thuật toán Random Forest nổi lên là thuật toán có hiệu suất vượt trội, đạt độ chính xác (Accuracy) 82% và chứng minh rằng Giới tính, Tuổi và Hạng hành khách (Passenger Class) là những đặc trưng chi phối và có ảnh hưởng lớn nhất trong quá trình dự đoán. Cuối cùng, những phát hiện này khẳng định hiệu quả của các phương pháp tổ hợp dựa trên cây (tree-based ensemble methods) và cung cấp bằng chứng thuyết phục về ứng dụng của các kỹ thuật khai phá dữ liệu trong dự đoán khả năng sống sót sau thảm họa.

Từ khóa: khai phá dữ liệu, phân tích dữ liệu khám phá, học máy, thảm họa hàng hải, phân loại, dự đoán khả năng sống sót, bộ dữ liệu Titanic, Random Forest.

1. Mở đầu

Vụ đắm tàu RMS Titanic đã trở thành một trường hợp điển hình mang lại giá trị to lớn cho lĩnh vực quản lý an toàn hàng hải. Như đã trình bày trong phần Tóm tắt, việc dự đoán khả năng sống sót của hành khách sau một thảm họa không chỉ mang ý nghĩa lịch sử sâu sắc mà còn thiết yếu trong các bài toán phân loại phức tạp, đặc biệt khi đối mặt với dữ liệu bị thiếu (missing data) và dữ liệu mất cân bằng (imbalanced data). Độ chính xác của mô hình dự đoán càng cao, thì khả năng liên hệ và áp dụng kết quả vào các đánh giá rủi ro trong đời sống thực tế càng mạnh mẽ.

Nghiên cứu này trình bày nhiệm vụ dự đoán khả năng sống sót dưới dạng một bài toán phân loại nhị phân có giám sát (supervised binary classification problem), sử dụng bộ dữ liệu Titanic được công bố rộng rãi trên nền tảng Kaggle. Các biến chính được sử dụng bao gồm các thuộc tính

tiềm năng (Giới tính, Tuổi), các đại diện cho yếu tố kinh tế-xã hội (Pclass, Fare) và các thuộc tính liên quan đến việc lên tàu (SibSp, Parch, Embarked).

Quy trình làm việc của chúng ta thể hiện việc áp dụng thành công các bước phân tích dữ liệu: phân tích dữ liệu khám phá (EDA), điền khuyết dữ liệu, mã hóa biến phân loại, và đánh giá so sánh các thuật toán học máy cổ điển, bao gồm Random Forest, Gradient Boosting, Logistic Regression và k-Nearest Neighbors (KNN). Việc đánh giá hiệu suất mô hình được thực hiện toàn diện, không chỉ tập trung vào độ chính xác (Accuracy) mà còn dựa trên các chỉ số chẩn đoán cụ thể cho từng lớp như Precision, Recall và F1-score, cùng với việc xây dựng ma trận nhầm lẫn (Confusion Matrix).

Những đóng góp chính của công trình này được xác định như sau:

- Xây dựng một quy trình cơ sở có thể tái tạo, nhằm làm nổi bật các yếu tố dự đoán vượt trội về khả năng sống sót trong dữ liệu hàng hải dạng bảng.
- Thực hiện so sánh thực nghiệm, chứng minh rằng các phương pháp tổ hợp dựa trên cây (Random Forest, Gradient Boosting,…) cho hiệu suất vượt trội hơn so với các mô hình cơ sở tuyến tính và dựa trên khoảng cách.
- Phân tích ngắn gọn về những thách thức trong tiền xử lý và các giới hạn của mô hình.

2. Cơ sở nghiên cứu và phương pháp nghiên cứu

2.1. Bộ dữ liệu thảm họa Titanic

Sự khám phá của chúng ta tập trung vào bộ dữ liệu "Titanic–Học máy từ Thảm họa", một thử thách phân loại nhị phân thú vị có nguồn gốc từ nền tảng Kaggle. Bộ dữ liệu được phân đoạn thành một tập huấn luyện (train.csv), bao gồm 891 mẫu quan sát, và một tập kiểm tra (test.csv), chứa 418 mẫu. Điều quan trọng là tập huấn luyện tích hợp biến mục tiêu "Survived" (0=Tử vong,1=Sống sót), biến này đóng vai trò là cơ sở cho việc huấn luyện và kiểm thử mô hình.

Bộ dữ liệu bao gồm 12 biến khác nhau minh họa các đặc điểm của hành khách, bao gồm hạng kinh tế-xã hội (Pclass), giới tính (Sex), tuổi (Age), người thân đi cùng (SibSp,Parch) và giá vé (Fare), bên cạnh các biến nhận dạng như Name, Ticket và Cabin như hình dưới đây. Tuy nhiên, việc khám phá dữ liệu ban đầu đã tiết lộ một thách thức nghiêm trọng: sự hiện diện của dữ liệu bị thiếu rất lớn trên các cột Age, Cabin và Embarked. Điều này đòi hỏi việc triển khai các quy trình làm sạch dữ liệu, điền khuyết và tiền xử lý nghiêm ngặt để đo lường chất lượng và sự phù hợp của dữ liệu cho việc triển khai mô hình học máy. Hình 2.1a, 2.1b và 2.2 cho thấy rõ ràng những gì chúng ta vừa thảo luận.

	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

hình 2.1a: 5 dòng đầu thông tin về bộ dữ liệu bao gồm các cột cùng với thông tin bộ train dataset

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	Q
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	S
2	894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	Q
3	895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN	S
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	NaN	S

hình 2.1b: 5 dòng đầu thông tin về bộ dữ liệu bao gồm các cột cùng với thông tin bộ test dataset

```
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
----------------------------------------
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 418 entries, 0 to 417
Data columns (total 11 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
...
 9   Cabin        91 non-null     object
 10  Embarked     418 non-null    object
dtypes: float64(2), int64(4), object(5)
```

hình 2.2: Tóm tắt về cấu trúc và kiểu dữ liệu của bộ train dataset

2.2. Tổng quan về vấn đề gặp phải

Mục tiêu của nghiên cứu này là giải quyết một bài toán phân loại nhị phân (binary classification issue) bằng cách dự đoán chính xác kết quả sống sót của hành khách trên tàu Titanic. Sự phân loại này được suy ra từ các thuộc tính nhân khẩu học và kinh tế-xã hội có sẵn trong bộ dữ liệu.

Tuân theo các phương pháp thực hành tốt nhất cho dữ liệu dạng bảng có cấu trúc, nghiên cứu này ưu tiên các thuật toán học máy truyền thống và tổ hợp. Sẽ có vài thuật toán không có hiệu

suất tốt do đặc điểm của chính nó và cả bộ dữ liệu này. Hơn hết, trong bộ dữ liệu thì có các giá trị nan/null, cần được xử lý nếu không sẽ ảnh hưởng tới mô hình cuối cùng.

2.4.2. Các phương thức đánh giá độ chính xác mô hình

Ngoài việc thao tác dữ liệu thông qua các kỹ thuật tiền xử lý, chúng ta nhận thấy rằng việc áp dụng các chiến lược cân bằng có thể giảm thiểu ảnh hưởng của sự mất cân bằng dữ liệu (data imbalance). Do đó, chúng ta đã thử nghiệm với các chỉ số đánh giá khác nhau để đo lường hiệu suất mô hình. Chúng ta sử dụng Ma trận Nhầm lẫn (Confusion Matrix) để tính toán tất cả các chỉ số, hiển thị số lượng True Positives (TP), False Positives (FP), True Negatives (TN) và False Negatives (FN) cho việc này. Tuy nhiên, sẽ thiếu cái nhìn tổng quan nếu quy trình này không có báo cáo phân loại (classification report).

Accuracy Score (Điểm Chính xác): Là chỉ số chính được sử dụng để đánh giá hiệu suất mô hình, được tính bằng tỷ lệ giữa số dự đoán đúng trên tổng số dự đoán:

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

Precision (Độ Chính xác): Đo lường tỷ lệ các trường hợp dương tính được dự đoán đúng trong số tất cả các trường hợp được dự đoán là dương tính:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall (Độ Nhạy/Thu Hồi): Đo lường tỷ lệ các trường hợp dương tính được dự đoán đúng trong số tất cả các trường hợp dương tính thực tế:$\text{Recall} = \frac{TP}{TP+FN}$

F1-Score: Là trung bình điều hòa của Precision và Recall, cung cấp thước đo cân bằng về hiệu suất mô hình:

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \times TP}{2 \times TP + FP + FN}$$

Hơn nữa, chúng ta cũng tích hợp thêm hai chỉ số bổ sung là trung bình macro (macro average) và trung bình có trọng số (weighted average). Trong một số tình huống, đặc biệt là trong trường hợp dữ liệu mất cân bằng, quá trình đánh giá sẽ bền vững và thực tế hơn khi trung bình macro và trung bình có trọng số thể hiện hiệu quả của chúng.

Macro Average (MAG): Tính giá trị trung bình cộng đơn giản của Precision, Recall và F1-score trên tất cả các lớp, đối xử với mỗi lớp như nhau bất kể kích thước của nó:

$$\text{Macro Average} = \frac{1}{N}\sum_{i=1}^{N} M_i$$

Weighted Average (WAG): Tính giá trị trung bình của Precision, Recall và F1-score trên tất cả các lớp, có trọng số theo số lượng mẫu (support) trong mỗi lớp:

$$\text{Weighted Average} = \frac{\sum_{i=1}^{N} w_i \times M_i}{\sum_{i=1}^{N} w_i}$$

, với (w_i) thể hiện trường hợp thực tế cho lớp (i)

3. Thí nghiệm và kết quả

3.1 Chuẩn bị thí nghiệm

Bộ Dữ liệu và Chia tách (Dataset and split):

Sử dụng train.csv với 891 mẫu..

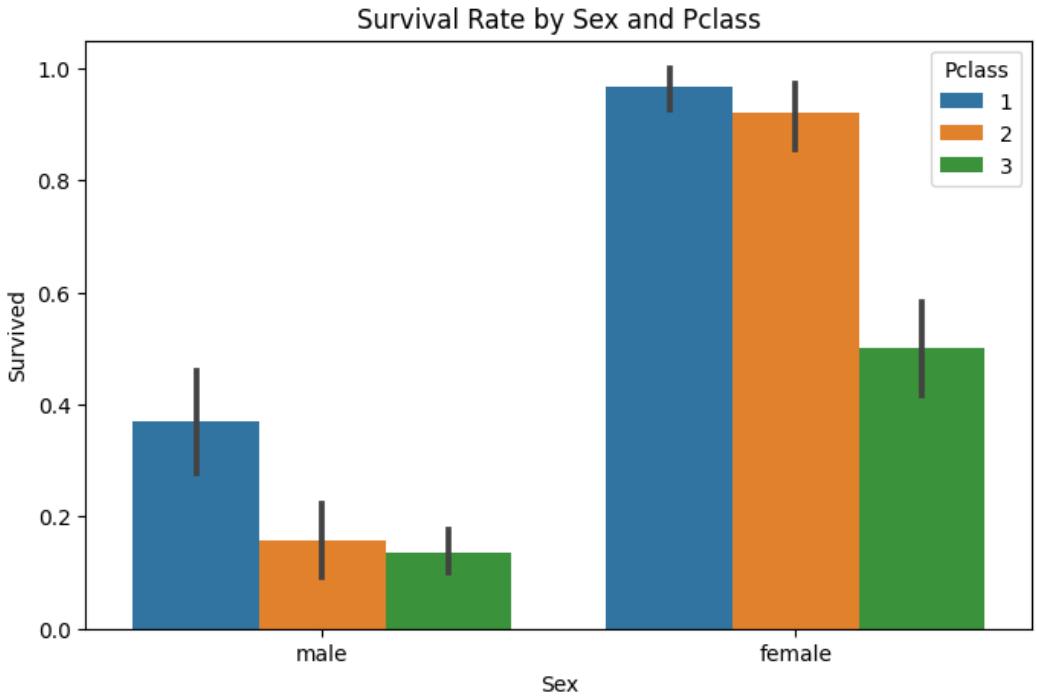
- Dữ liệu được chia một lần thành các tập huấn luyện và kiểm với tỷ lệ 80/20 bằng cách sử dụng train_test_split.
- Không thực hiện kiểm định chéo (cross-validation) hoặc điều chỉnh siêu tham số (hyperparameter tuning).

Tiền xử lý (Preprocessing):

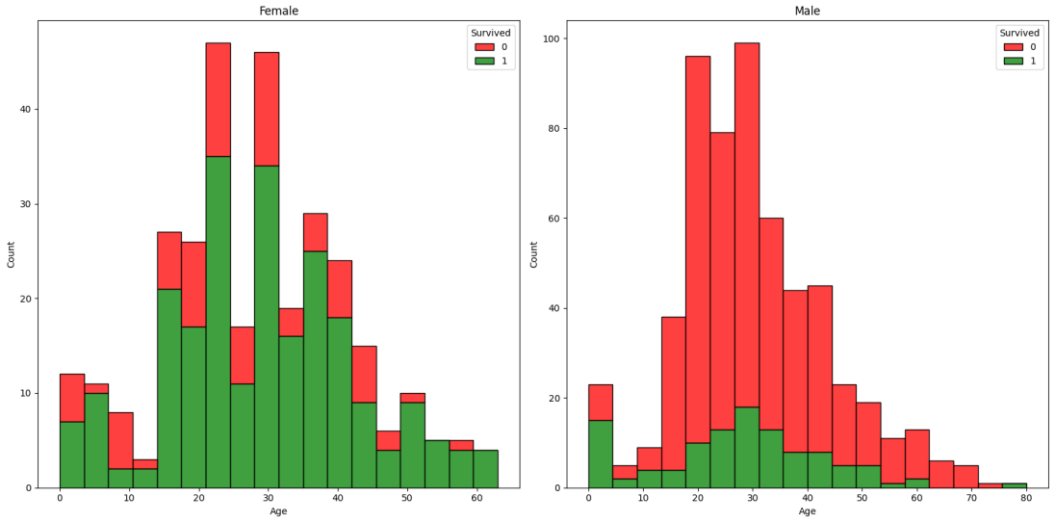
- Điền khuyết (Imputation): Điền các giá trị Tuổi (Age) và Giá vé (Fare) bị thiếu bằng trung vị (median); điền Cảng lên tàu (Embarked) bị thiếu bằng mode (mode).
- Mã hóa (Encoding): Áp dụng Label Encoding cho Giới tính (Sex) và Cảng lên tàu (Embarked). Không sử dụng chuẩn hóa đặc trưng (feature scaling), One-Hot Encoding, hoặc các đặc trưng được kỹ thuật hóa (engineered features) (ví dụ: Danh xưng (Title), Quy mô gia đình (Family Size), Đi một mình (IsAlone).
- Đặc trưng được sử dụng: Pclass, Sex, Age, Fare, SibSp, Parch, Embarked.
- Các Mô hình được đánh giá: Random Forest, Gradient Boosting, Logistic Regression, và K-Nearest Neighbors. Các mô hình được huấn luyện trên phần chia tách huấn luyện và được đánh giá trên phần kiểm tra được giữ lại (hold-out test split).

Các chỉ số đánh giá chính:

- Chỉ số chính là Độ chính xác (Accuracy) trên tập giữ lại (hold-out set).
- Ngoài ra, ma trận nhầm lẫn (confusion matrices) và các báo cáo phân loại (classification reports) như precision, recall, F1 cũng được tạo ra.
- Theo định nghĩa của thiết lập thực nghiệm bị ràng buộc, một bước Kiểm chứng Dữ liệu Trực quan (Visual Data Validation) đã được thực hiện.
- Các hình 3.1a và 3.1b trình bày một số kết quả nổi bật của Phân tích Dữ liệu Khám phá (EDA), nhằm xác nhận rằng các đặc trưng cơ bản được chọn-đặc biệt là Pclass, Sex và Age-sở hữu khả năng phân biệt đủ mạnh.



hình 3.1a: Biểu đồ cột ba thể hiện tỷ lệ sống sót được thể hiện qua giới tính và hạng vé



hình 3.1b: Biểu đồ cột chồng thể hiện tổng số người sống và chết ở cùng các độ tuổi khác nhau, chia ra nam và nữ

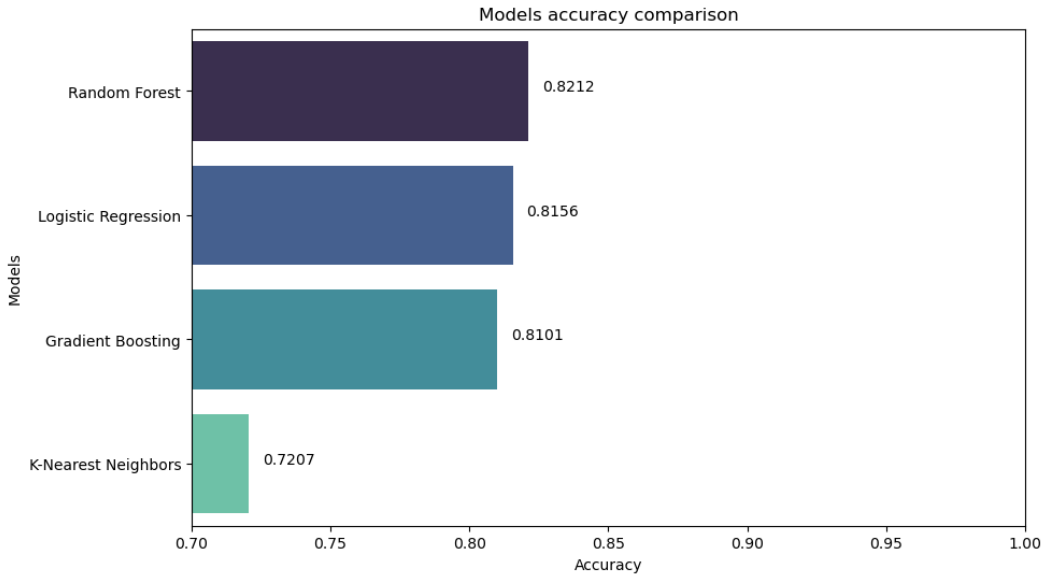
Như hình 3.1a và 3.1b thể hiện, rõ ràng tỷ lệ sống sót khác biệt đáng kể theo giới tính và độ tuổi. Phụ nữ, đặc biệt là những người trẻ tuổi và trung niên, có tỷ lệ sống sót cao hơn nhiều so với nam giới. Trẻ em cũng có cơ hội sống sót cao hơn so với người lớn. Ngoài ra, hành khách ở các

hạng cao hơn (Pclass 1 và 2) có nhiều khả năng sống sót hơn, đặc biệt là ở phụ nữ. Những mô hình này cho thấy Giới tính, Độ tuổi và Pclass là những yếu tố chính ảnh hưởng đến tỷ lệ sống sót.

Trước khi huấn luyện các mô hình, chúng ta có thể đưa ra giả thuyết rằng các mô hình phi tuyến tính như Random Forest hoặc Gradient Boosting có thể sẽ hoạt động tốt hơn các mô hình tuyến tính như Logistic Regression, vì chúng có thể nắm bắt tốt hơn các tương tác phức tạp giữa các đặc điểm này.

3.2. Kết quả

Sau khi đã chuẩn bị hết các bước, ta áp dụng các nghiệp vụ lập trình. Hình 3.2 và bảng 3.1 sẽ là kết quả của các mô hình khi đứng chung với nhau:



hình 3.2: Biểu đồ cột ngang thể hiện sự so sánh độ chính xác (accuracy) của các mô hình huấn luyện

Và bảng 3.1 thể hiện sự so sánh các chỉ số đánh giá kết quả của tất cả các tham số đánh giá:

Bảng 3.1: Bảng so sánh các kết quả đánh giá của các mô hình sau khi huấn luyện

Model	**Accuracy**	**Precision (macro)**	**Recall (macro)**	**F1-score (macro)**
Random Forest	**0.8212**	0.82	0.81	0.81
Gradient Boosting	0.8101	0.81	0.79	0.80
Logistic Regression	0.8156	0.81	**0.82**	**0.81**
K-Nearest Neighbors (KNN)	0.7207	0.72	0.69	0.70

Từ các kết quả, cả ba mô hình tổ hợp và mô hình tuyến tính như Random Forest, Gradient Boosting, và Logistic Regression-đều đạt hiệu suất tương đối tương đương và mạnh mẽ, với độ chính xác accuracy và điểm F1 dao động quanh mức 0.81-0.82 Trong số đó, Random Forest đạt

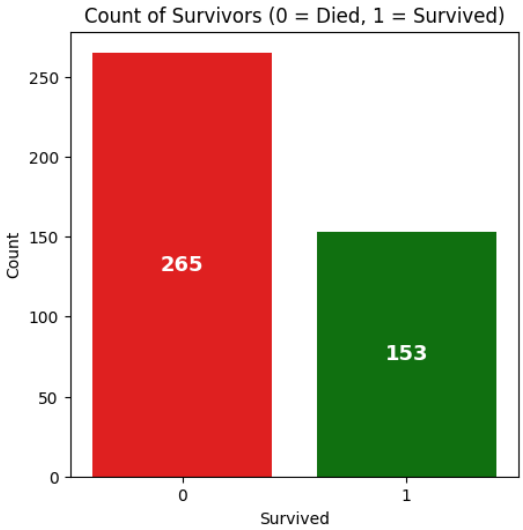
hiệu suất tốt nhất xét về tổng thể, thể hiện sự đánh đổi cân bằng giữa precision, recall và điểm F1.

Điều này xác nhận rằng các phương pháp tổ hợp dựa trên cây (ensemble tree-based methods) có thể nắm bắt hiệu quả các mô hình phi tuyến tính và tương tác đặc trưng (chẳng hạn như giữa Giới tính, Tuổi và Hạng hành khách) có mặt trong bộ dữ liệu Titanic.

Ngược lại, K-Nearest Neighbors (KNN) có hiệu suất kém hơn đáng kể, với độ chính xác chỉ 0.72 và các giá trị recall và F1 thấp hơn. Hiệu suất thấp này là điều được dự đoán bởi vì nó dựa vào các chỉ số khoảng cách trong không gian đặc trưng, vốn hoạt động kém khi các đặc trưng là phân loại hoặc có các thang đo khác nhau (ví dụ: Sex, Pclass, Embarked). Hơn nữa, KNN nhạy cảm với nhiễu và sự mất cân bằng dữ liệu, khiến nó kém hiệu quả đối với dữ liệu dạng bảng có cấu trúc như Titanic, nơi các ranh giới quyết định không trơn tru mà mang tính rời rạc và dựa trên quy tắc-các mô hình mà mô hình dựa trên cây nắm bắt tốt hơn nhiều.

3.3 Kết quả dự đoán của mô hình tốt nhất

Mô hình Random Forest được lựa chọn đã đạt độ chính xác 82%, được so sánh với các mô hình khác qua hình 3.2. Sau khi đã chọn ra mô hình có hiệu suất tốt nhất, ta sẽ dùng chính nó để dự đoán những hành khách không có thông tin về cột sống sót, hình 3.3 dưới đây thể hiện tổng người sống và chết-kết quả của mô hình.



hình 3.3: Tổng số người sống sót và thiệt mạng ở vụ chìm tàu Titanic

Nhưng ta hãy tự đặt câu hỏi xem, liệu các phương thức đánh giá mô hình có chính xác so với kết quả thực tế? Phần sau đây sẽ giải đáp thắc mắc.

3.3. Phân tích sự khác biệt về độ chính xác giữa dự đoán và thực tế

3.3.1. Hướng tiếp cận

Để đánh giá hiệu suất của mô hình trên tập kiểm tra, ta cần đối chiếu các kết quả dự đoán với trạng thái sống sót lịch sử đã được xác minh của từng hành khách. Dữ liệu thực tế cho sự so sánh này đã được thu thập tỉ mỉ và xác minh bằng cách sử dụng Encyclopedia Titanica, cơ quan đăng ký lịch sử

xác định về hành khách Titanic, và kết quả này cũng đã được đánh giá chuẩn (Benchmark) trên thử thách Kaggle – nơi cung cấp bộ dữ liệu [1], với kết quả thực tế là kết quả từ chính trang kaggle thư thách Titanic[3].

3.3.2. Kết quả và giải thích

Kết quả đánh giá mô hình phân loại hành khách Titanic cho thấy sự khác biệt đáng kể giữa hiệu suất nội bộ và hiệu suất ngoại suy. Cụ thể, mô hình đạt Độ chính xác cục bộ (Local Accuracy) là 82% khi được đánh giá trên tập dữ liệu kiểm tra nội bộ đã biết, nhưng chỉ đạt 75% trên Bảng xếp hạng Công khai Kaggle (Kaggle Public Leaderboard). Sự sụt giảm 7% này là minh chứng rõ ràng cho hiện tượng Quá khớp (Overfitting).

Quá khớp xảy ra khi mô hình học quá sát các đặc điểm và nhiễu ngẫu nhiên của tập dữ liệu huấn luyện và tập kiểm tra cục bộ, khiến nó mất đi khả năng tổng quát hóa (Generalization). Thay vì nắm bắt được các mẫu cơ bản của dữ liệu, mô hình đã trở thành một "công cụ ghi nhớ" hiệu suất cao. Khi đối diện với dữ liệu chưa từng thấy (dữ liệu bí mật của Kaggle dùng để chấm điểm), mô hình không thể áp dụng các quy tắc đã học, dẫn đến hiệu suất giảm mạnh.

Hiện tượng này được làm trầm trọng thêm bởi việc thiếu các kỹ thuật chống quá khớp và sử dụng mã hóa không phù hợp. Cụ thể:

- Sử dụng Label Encoding trên các biến phân loại danh nghĩa đã áp đặt giả định thứ bậc sai lệch lên dữ liệu, làm sai lệch các mối quan hệ và khiến mô hình học các quy tắc quá cụ thể.
- Không sử dụng Kiểm định chéo (Cross-Validation) trong quá trình huấn luyện đã ngăn cản việc có được ước tính hiệu suất thực tế, dẫn đến việc đánh giá quá cao khả năng của mô hình.
- Do đó, độ chính xác 82% là một con số bị thổi phồng, trong khi 75% phản ánh hiệu suất thực tế của mô hình trên dữ liệu mới. Việc cải thiện mô hình cần tập trung vào việc tăng cường khả năng tổng quát hóa, ưu tiên các phương pháp như One-Hot Encoding và điều chỉnh tham số mô hình để giảm độ phức tạp.

3.3. Hướng giải quyết

Từ các vấn đê tồn đọng và thiếu sót căn bản của bài, ta sẽ có các cách giải quyết cũng như mở rộng sau hơn:

- Sử dụng kiểm định chéo (Cross-validation) để chọn ra tham số tối ưu.
- Sử dụng các phương pháp xử lý cho dữ liệu nan/null khác thay vì điền các giá trị trung bình/trung vị
- Sử dụng kỹ thuật “feature engineering” để tạo ra các cột mới có từ các cột có sẵn để tạo ra thuộc tính có sức mạnh và giúp mô hình học tốt hơn, đã được chính minh là có hiệu quả.

4. Kết luận

Việc đánh giá nhiều mô hình cho dự đoán khả năng sống sót trên tàu Titanic đã mang lại kết quả rõ ràng và đáng tin cậy, xác nhận rằng tiền xử lý nhẹ nhàng kết hợp với các phương pháp tổ hợp vẫn hoạt động hiệu quả trên dữ liệu dạng bảng. Những phát hiện của chúng ta cho thấy các thuật toán

học máy truyền thống, ngay cả khi không có điều chỉnh mở rộng, vẫn giữ vững độ mạnh và tính thực tiễn khi được cung cấp các đặc trưng đã được điền khuyết thích hợp và mã hóa đơn giản. Tuy nhiên, để mô hình được chính xác hơn nữa thì không chỉ dừng lại ở các bước cơ bản ta cũng cần tìm hiểu và tranh luận sâu thêm để tìm ra hướng xử lý từ đầu hiệu quả và có tổ chức hơn.

Hiệu quả của Tiền xử lý: Chúng ta đã sử dụng làm sạch dữ liệu cơ bản với điền khuyết bằng trung vị và label encoding cho các biến phân loại. Các bước này đã giảm thiểu ảnh hưởng của dữ liệu bị thiếu và đảm bảo biểu diễn đặc trưng ổn định, mang lại sự cải thiện nhất quán về độ chính xác của mô hình. Nhưng cũng chính vì thế nó cũng làm mô hình học được các quy tắc mà sẽ gây overfitting

Ưu thế của Phương pháp tổ hợp: Trong số bốn thuật toán được đánh giá, Random Forest đã đạt độ chính xác cao nhất là 82%. Điều này chứng minh rằng các mô hình tổ hợp dựa trên cây có thể nắm bắt hiệu quả các mối quan hệ phi tuyến tính trong bộ dữ liệu Titanic, vượt trội hơn các phương pháp đơn giản hơn như Logistic Regression và K-Nearest Neighbors trong các điều kiện giống hệt nhau.

Tóm lại, tiền xử lý đơn giản kết hợp với các mô hình tổ hợp mạnh mẽ mang lại kết quả khả quan trong nhiệm vụ này. Công việc trong tương lai nên tập trung vào tối ưu hóa siêu tham số có hệ thống và kỹ thuật đặc trưng để nâng cao hơn nữa hiệu suất.

5. Lời cảm ơn

Bài này ban đầu được phát triển như một phần của môn Khai phá Dữ liệu tại Đại học Sài Gòn. Chúng em xin chân thành cảm ơn thầy Đỗ Như Tài đã hướng dẫn chi tiết cũng như là cung cấp tài liệu tương tự để có thể tham khảo.

Tham Khảo

[1] Kaggle, “Titanic – Machine Learning from Disaster,” in Kaggle Competitions, 2012. [Online]. Available: https://www.kaggle.com/competitions/Titanic

[2] Patil, A., Singh, D.: Predicting survival on the Titanic using machine learning techniques. International Journal of Innovative Research in Computer and Communication Engineering, 5(4), 6429–6435 (2017).

[3] Encyclopedia Titanica. (n.d.). Retrieved from https://www.encyclopedia-Titanica.org/