Feature Selection

Lựa chọn features

PROTON X

PROTONX

PROTON X

Data-centric vs Model-centric

Data-centric vs Model-centric

	Steel defect detection	Solar panel	Surface inspection
Baseline	76.2%	75.68%	85.05%
Model-centric	+0% (76.2%)	+0.04% (75.72%)	+0.00% (85.05%)
Data-centric	+16.9% (93.1%)	+3.06% (78.74%)	+0.4% (85.45%)

Model-centric Giữ nguyên dữ liệu, phát triển model để tăng độ chính xác trên dữ liệu

Data-centric Sử dụng mô hình cố định và sử dụng các công cụ để **tăng cường chất lượng dữ liệu**

Dữ liệu có cấu trúc và không cấu trúc

Structured vs unstructured data

Structured Data

PassengerId	Survived	Pclass
1	0	3
2	1	1
3	1	3
4	1	1
5	0	3

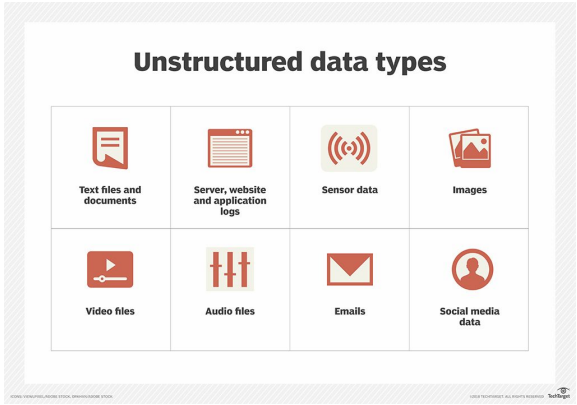
Titanic Dataset

Với structured data, chúng ta có lượng dữ liệu nhỏ kèm theo một số lượng thuộc tính cố định và nhiều yếu tố cố định khác đi kèm. Ví dụ như số lượng khách hàng cố định, sản phẩm cố định

Thực tế cho thấy các mô hình học sâu không có nhiều đột biến trên dạng dữ liệu này.

Khi làm việc với dữ liệu này nên đi theo hướng
- **Data Centric:** Thêm, xóa, sửa đặc trưng để tăng chất lượng dữ liệu
- Sử dụng các mô hình ML truyền thông
- Các mô hình nổi tiếng: XGBoost, Lightgbm.

Unstructured Data



Với unstructured data, chúng ta có thể áp dụng các kỹ thuật học sâu tiên tiến:
- Tăng cường dữ liệu (Data Augmentation)
- Transfer Learning
- Các mô hình học sâu
- etc

https://searchbusinessanalytics.techtarget.com/definition/unstructured-data

PROTON X

Bộ dữ liệu Titanic

Titanic Dataset

Hạng ghế

1: Thương gia

2: Phổ thông

3: Phổ thông tiết kiệm

Số lượng vợ chồng/anh chị em ruột trên tàu

Số lượng cha mẹ/con trên tàu

C: Cherbourg

Q: Queenstown

S: Southampton

Cảng lên tàu

	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

Khả năng sống sót

Mã Vé

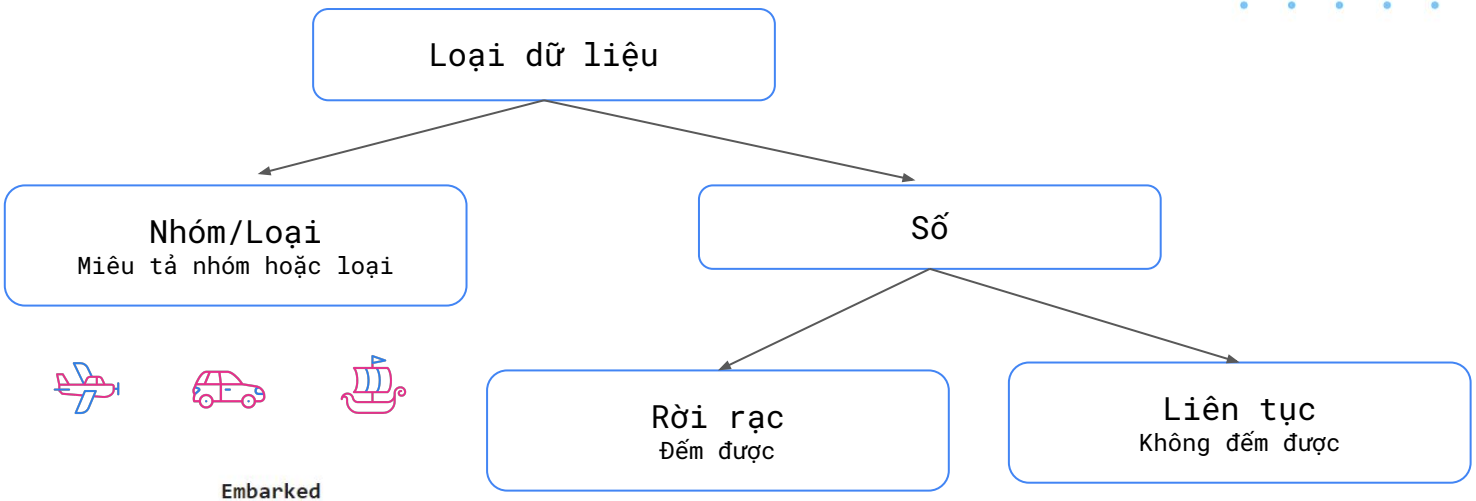
Buồng

Chứa cả ký tự và số

Chi tiết về bộ dữ liệu

Đặc trưng số và đặc trưng phân loại

Numerical feature vs Categorical feature



Có duy nhất 3 cổng C, S và Q để lên tàu

Embarked
S
C
S
S
S

Số lượng vợ chồng/ anh chị em là đếm được

SibSp
1
1
0
1
0

PROTON X

Đặc trưng số

Numerical feature

```
train_df.describe(percentiles=[.61, .62])
```

Tỉ lệ sống sót: ≈ 38%

	PassengerId	Survived
count	891.000000	891.000000
mean	446.000000	0.383838
std	257.353842	0.486592
min	1.000000	0.000000
50%	446.000000	0.000000
61%	543.900000	0.000000
62%	552.800000	1.000000
max	891.000000	1.000000

```
train_df.describe(percentiles=[.76, .77])
```

≈ 76% không mang theo cha mẹ/con cái

	Parch
count	891.000000
mean	0.381594
std	0.806057
min	0.000000
50%	0.000000
76%	0.000000
77%	1.000000
max	6.000000

```
train_df.describe(percentiles=[.90, .99])
```

Tuổi từ 65 đến 80 chiếm rất ít (< 1%)

	Age
count	714.000000
mean	29.699118
std	14.526497
min	0.420000
50%	28.000000
80%	41.000000
90%	50.000000
99%	65.870000
max	80.000000

PROTON X

Đặc trưng Nhóm/Loại

Categorical feature

Không có ai trùng tên

```
# select pandas categorical columns
train_df.describe(include=['O'])
```

	Name	Sex	Ticket	Cabin	Embarked
count	891	891	891	204	889
unique	891	2	681	147	3
top	Perkin, Mr. John Henry	male	1601	G6	S
freq	1	577	7	4	644

Nhiều khách lên tàu thông qua cổng S nhất

Có 4 người ở Cabin G6

PROTON X

Giả định

Assumptions

1. Người già và trẻ em sẽ được ưu tiên cứu hộ trước
2. Người giàu sẽ được ưu tiên cứu trước
3. Ai mới kiếm được bạn gái trên tàu sẽ chết trước :(
4. Phụ nữ luôn được ưu tiên

Giả định được đưa ra **thông qua những phân tích ban đầu** và sẽ được điều chỉnh lại khi sau khi phân tích chi tiết

PROTON X

Lựa chọn đặc trưng

Feature Selection

Phân tích

VIP

Hạng thương gia có **khả năng sống sót cao hơn** các hạng còn lại

```
[47] train_df[['Pclass', 'Survived']].groupby(['Pclass'], as_index=False) \
     .mean().sort_values(by='Survived', ascending=False)
```

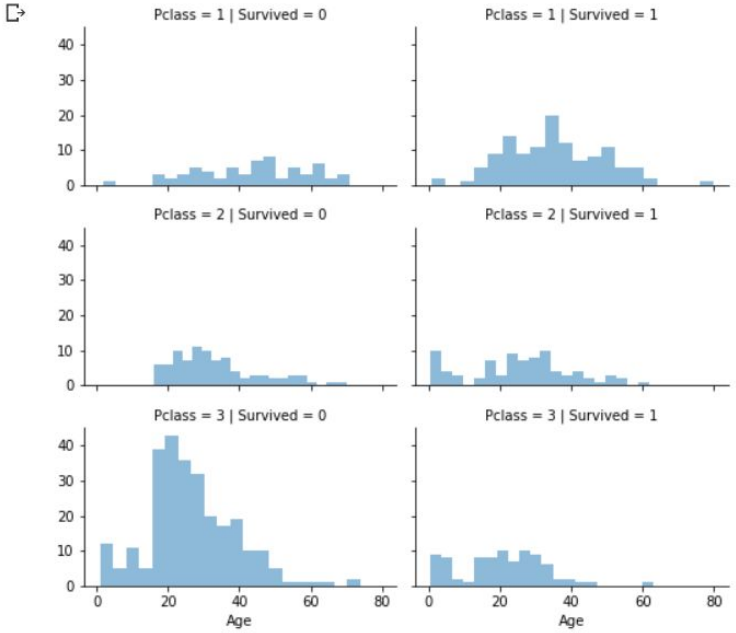
	Pclass	Survived
0	1	0.629630
1	2	0.472826
2	3	0.242363

PROTON X

Lựa chọn đặc trưng

Feature Selection

```
# grid = sns.FacetGrid(train_df, col='Pclass', hue='Survived')
grid = sns.FacetGrid(train_df, col='Survived', row='Pclass', height=2.2, aspect=1.6)
grid.map(plt.hist, 'Age', alpha=.5, bins=20)
grid.add_legend();
```



Phân tích

- Hầu hết hành khách hạng 1 sống sót
- Khách hàng hạng 3 nhiều nhất và **hầu hết là không sống sót**
- Trẻ nhỏ hạng 2 **hầu hết là sống sót**

Quyết định

- Thêm đặc trưng Pclass vào mô hình

Chú ý: Tương quan giữa đặc trưng số và loại (đang ở dạng số)

PROTON X

Lựa chọn đặc trưng

Feature Selection

Phân tích

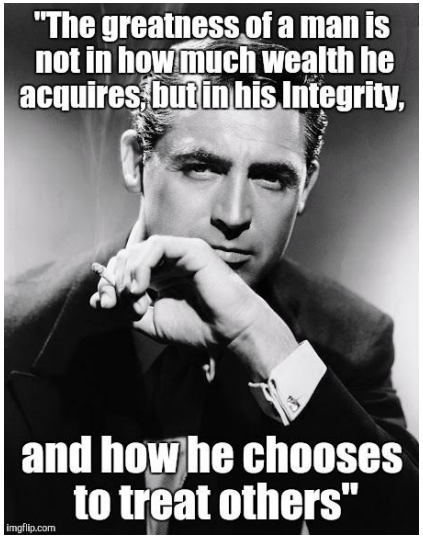
Với sự ga lăng của những gentleman thì phụ nữ có khả năng sống sót hơn rất nhiều

```
[48] train_df[["Sex", "Survived"]].groupby(['Sex'], as_index=False) \
     .mean().sort_values(by='Survived', ascending=False)
```

	Sex	Survived
0	female	0.742038
1	male	0.188908

Quyết định

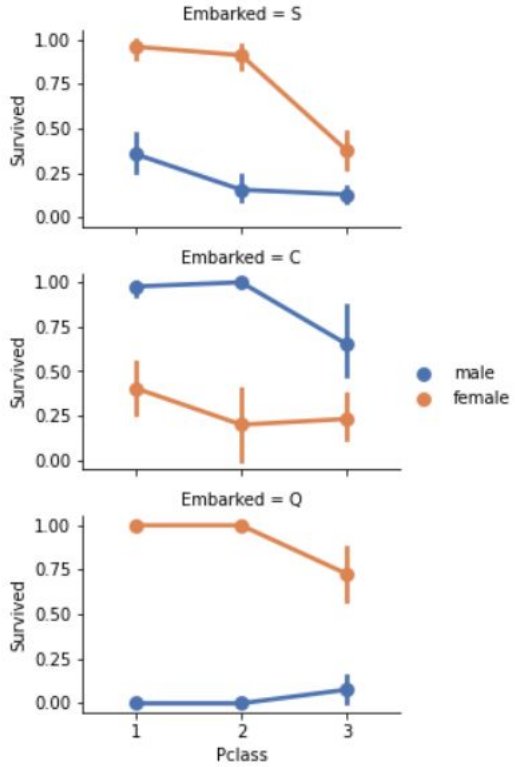
Đặc trưng **Giới tính** sẽ được thêm vào mô hình



PROTON X

Lựa chọn đặc trưng

Feature Selection




```
# grid = sns.FacetGrid(train_df, col='Embarked')
grid = sns.FacetGrid(train_df, row='Embarked', height=2.2, aspect=1.6)
grid.map(sns.pointplot, 'Pclass', 'Survived', 'Sex', palette='deep')
grid.add_legend()
```

Phân tích

- Ở cửa lên (Embarked) C nam có tỷ lệ sống sót cao hơn nữ.
- Tỷ lệ nam sống sót với Pclass = 3 cao hơn Pclass = 2 khi ở cổng Embarked = Q
- Các cổng đón khách Pclass = 3 có tỉ lệ sống sót rất khác nhau

Quyết định

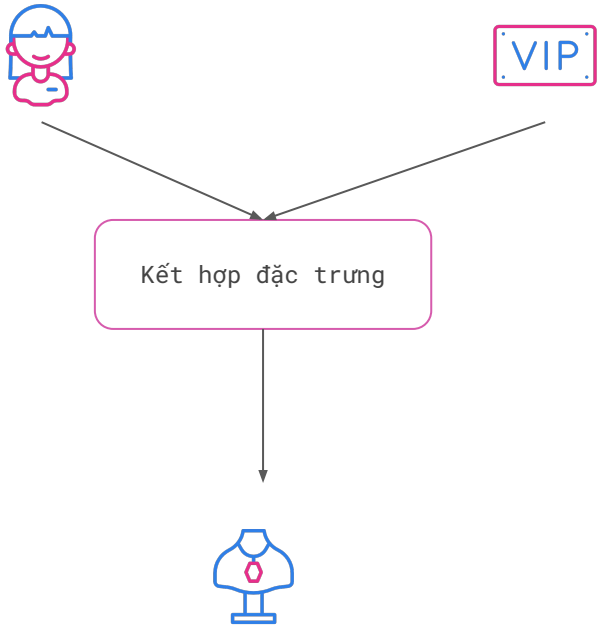
- Bổ sung các giá trị Embarked thiếu
- Thêm đặc trưng Embarked vào mô hình

Chú ý: Tương quan giữa các đặc trưng loại với nhau

PROTON X

Kết hợp đặc trưng

Feature Combination



Kết hợp này chỉ mang tính chất **minh họa**

PROTON X

Lựa chọn đặc trưng

Feature Selection

Phân tích

Liệu rằng có sự tương quan chặt chẽ giữa
số lượng người thân trên tàu và khả năng sống sót ???

```
[50] train_df[["SibSp", "Survived"]].groupby(['SibSp'], as_index=False). \
     mean().sort_values(by='Survived', ascending=False)
```

	SibSp	Survived
1	1	0.535885
2	2	0.464286
0	0	0.345395
3	3	0.250000
4	4	0.166667
5	5	0.000000
6	8	0.000000

```
train_df[["Parch", "Survived"]].groupby(['Parch'], as_index=False). \
mean().sort_values(by='Survived', ascending=False)
```

	Parch	Survived
3	3	0.600000
1	1	0.550847
2	2	0.500000
0	0	0.343658
5	5	0.200000
4	4	0.000000
6	6	0.000000

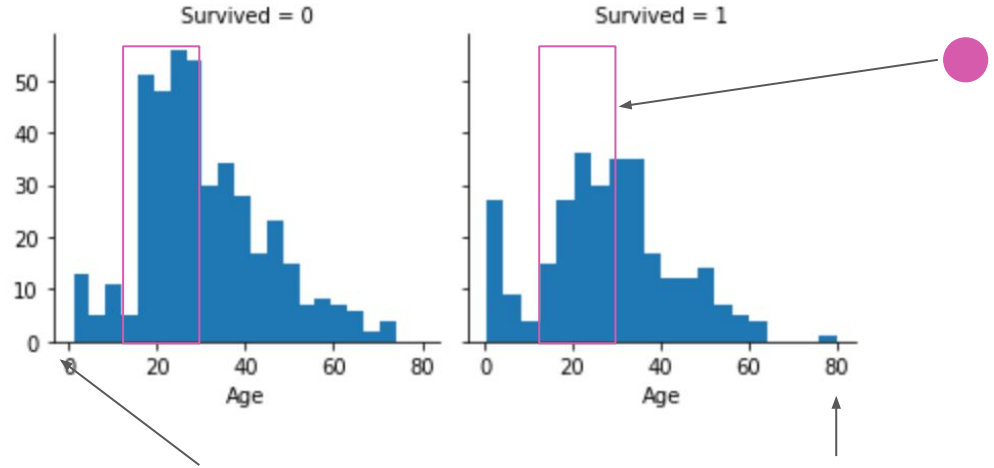
Mối quan hệ này **vẫn chưa thực sự rõ ràng**

Lựa chọn đặc trưng

Feature Selection

```
import seaborn as sns
g = sns.FacetGrid(train_df, col='Survived')
g.map(plt.hist, 'Age', bins=20)
```

```
<seaborn.axisgrid.FacetGrid at 0x7fd2552bd3d0>
```



Trẻ nhỏ (<= 4 tuổi)
có khả năng sống sót cao

Người lớn tuổi nhất
đã sống sót

Phân tích

Khoảng **15 - 30 tuổi**
có khả năng chết cao
hơn sống

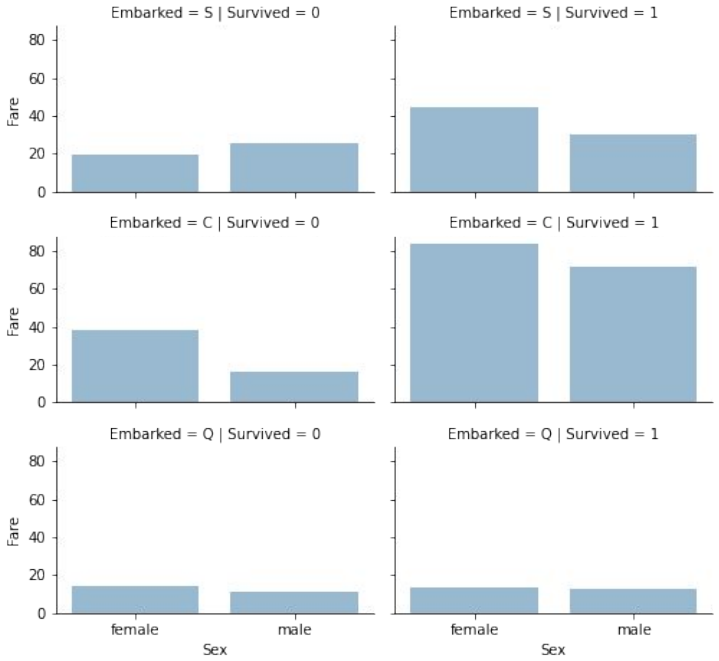
Quyết định

- Thêm đặc trưng tuổi vào mô hình
- Phân tuổi **thành các nhóm**

PROTON X

Lựa chọn đặc trưng

Feature Selection



Chú ý: Tương quan giữa đặc trưng số và loại (đang ở dạng chuỗi)

```
# grid = sns.FacetGrid(train_df, col='Embarked', hue='Survived', palette={0: 'k', 1: 'w'})
grid = sns.FacetGrid(train_df, row='Embarked', col='Survived', size=2.2, aspect=1.6)
grid.map(sns.barplot, 'Sex', 'Fare', alpha=.5, ci=None)
grid.add_legend()
```

Phân tích

- Người trả giá vé cao hơn có tỉ lệ sống sót cao hơn
- Có mối tương quan giữa cổng vào (Embarked) với tỉ lệ sống sót

Quyết định

- Tạo khoảng cho đặc trưng Fare
- Thêm đặc trưng Fare vào mô hình

PROTON X

Xóa đặc trưng

Feature Dropping

```
# select pandas categorical columns
train_df.describe(include=['O'])
```

	Name	Sex	Ticket	Cabin	Embarked
count	891	891	891	204	889
unique	891	2	681	147	3
top	Perkin, Mr. John Henry	male	1601	G6	S
freq	1	577	7	4	644

Phân tích

- 22% giá trị vé bị trùng
- Thiếu rất nhiều giá trị Cabin

Quyết định

- Không sử dụng đặc trưng Ticket cho mô hình
- Không sử dụng đặc trưng Cabin cho mô hình

Practice

Thực hành

PROTON X

PROTONX