Dự đoán sự hài lòng của khách hàng ngân hàng (LR, SVM, RF)

Người trình bày: Nguyễn Văn Đạt

Giảng viên hướng dẫn: PGS.TS Nguyễn Văn Hậu

Date: December 3rd, 2025

Contents

Contents

I. Tổng quan bài toán

- Trong lĩnh vực ngân hàng việc giữ chân khách hàng thường ít tốn kém hơn so với việc thu hút khách hàng mới.
- Bài toán Bank Churn là một bài toán **Binary Classification**, nhằm mục đích dự đoán xem một khách hàng có khả năng **rời bỏ** ngân hàng hay không.

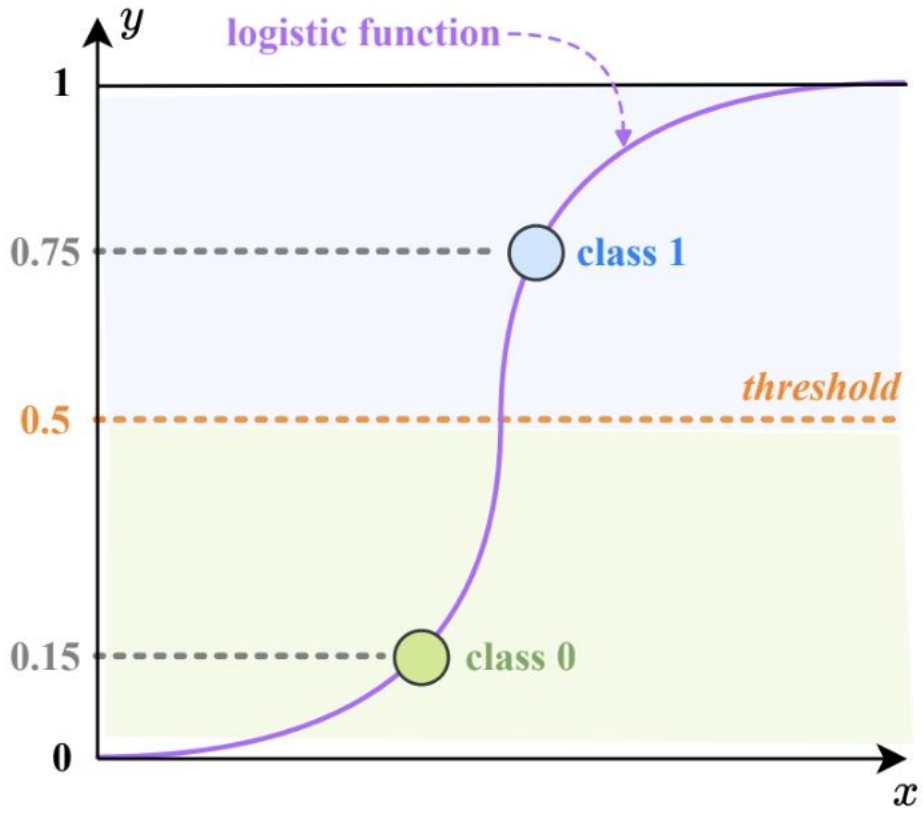
Contents

II. Mô hình học máy

- Các mô hình sử dụng trong bài toán:
  1. Logistic Regression
  2. Support Vector Machine
  3. Random Forest

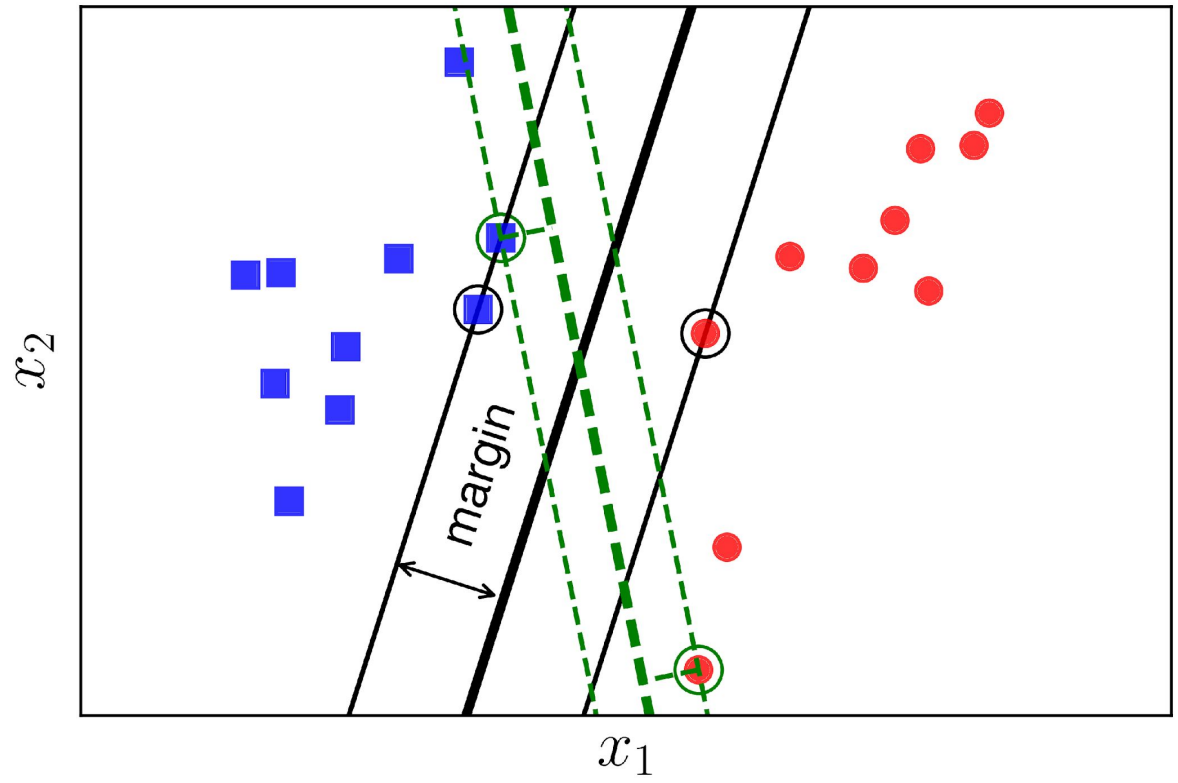
Logistic Regression

Định nghĩa: Logistic Regression là một mô hình supervised learning trong học máy dùng để giải quyết bài toán phân loại nhị phân.



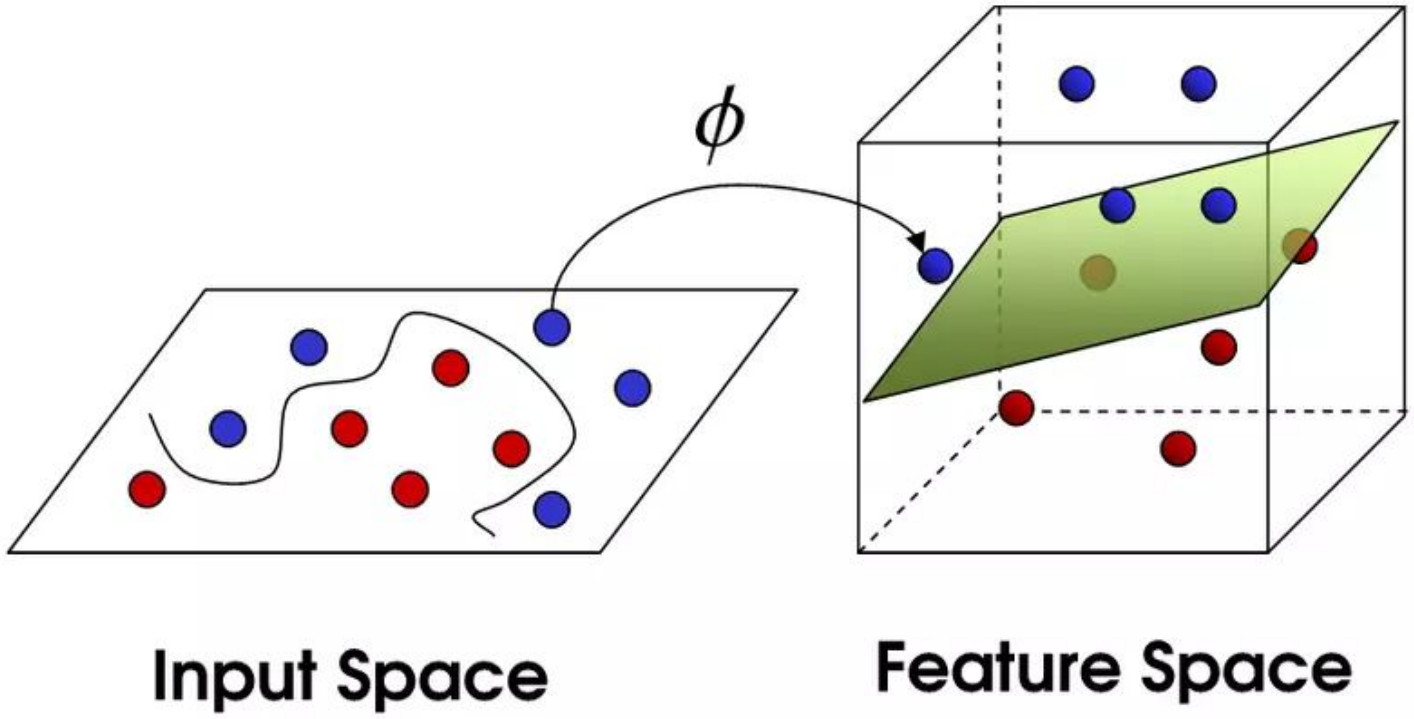
Support Vector Machine

Định nghĩa: SVM là một thuật toán học có giám sát (supervised learning), chủ yếu được sử dụng cho phân loại (classification)



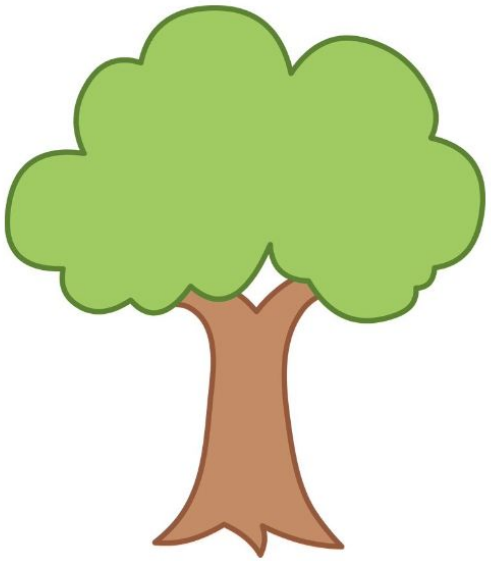
Support Vector Machine

Một yếu tố làm nên hiệu quả của SVM đó là việc sử dụng Kernel function khiến cho các phương pháp chuyển không gian trở nên linh hoạt hơn.



Random Forest

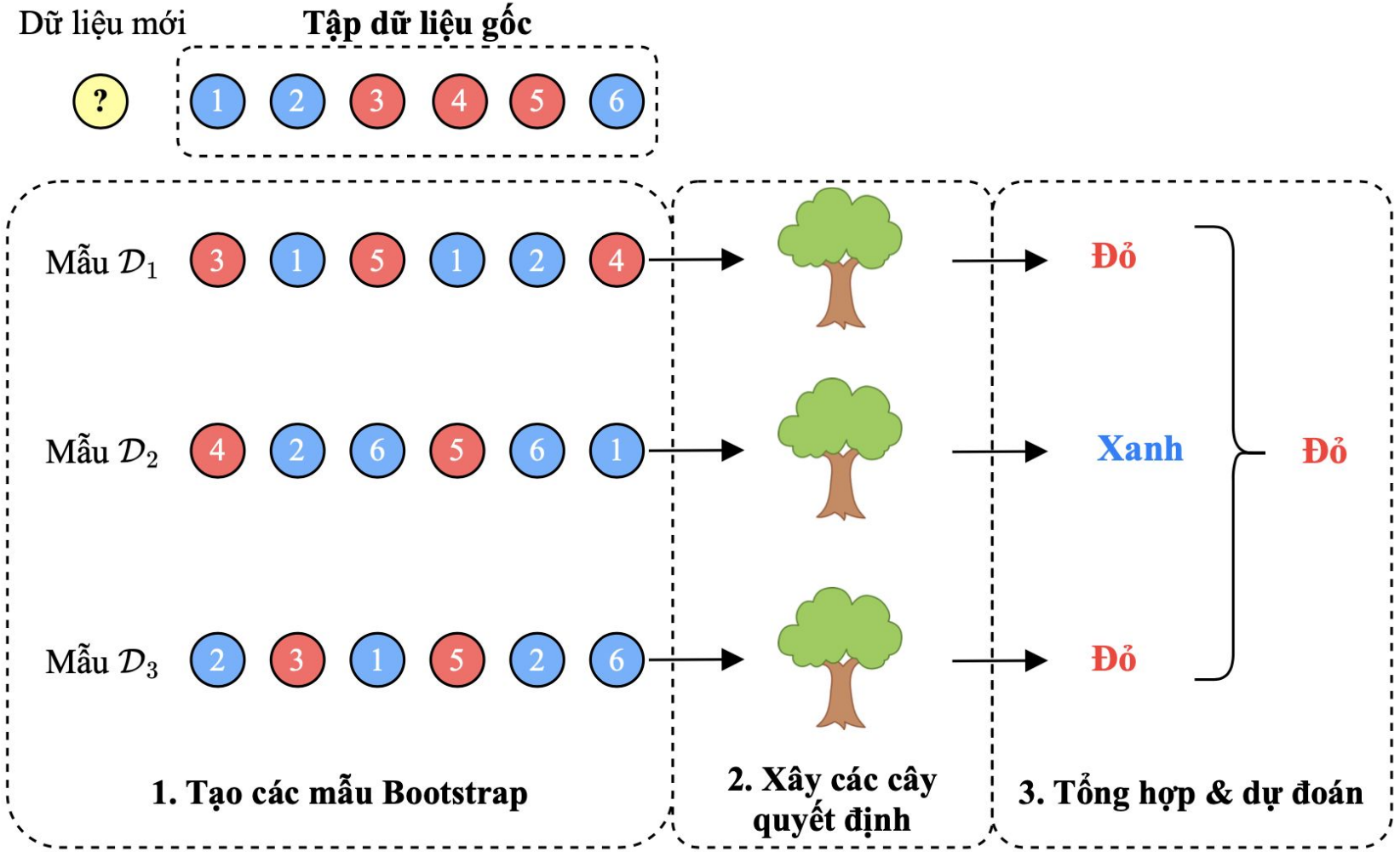
Định nghĩa: Random Forest là một thuật toán học tập hợp (Ensemble Learning), được sử dụng cho cả bài toán phân loại (classification) và hồi quy (regression).



Deicision Tree

Random Forest

Random Forest



Metrics đánh giá

- Accuracy
- Precision
- Recall
- F1-score

Contents

Mô tả dữ liệu

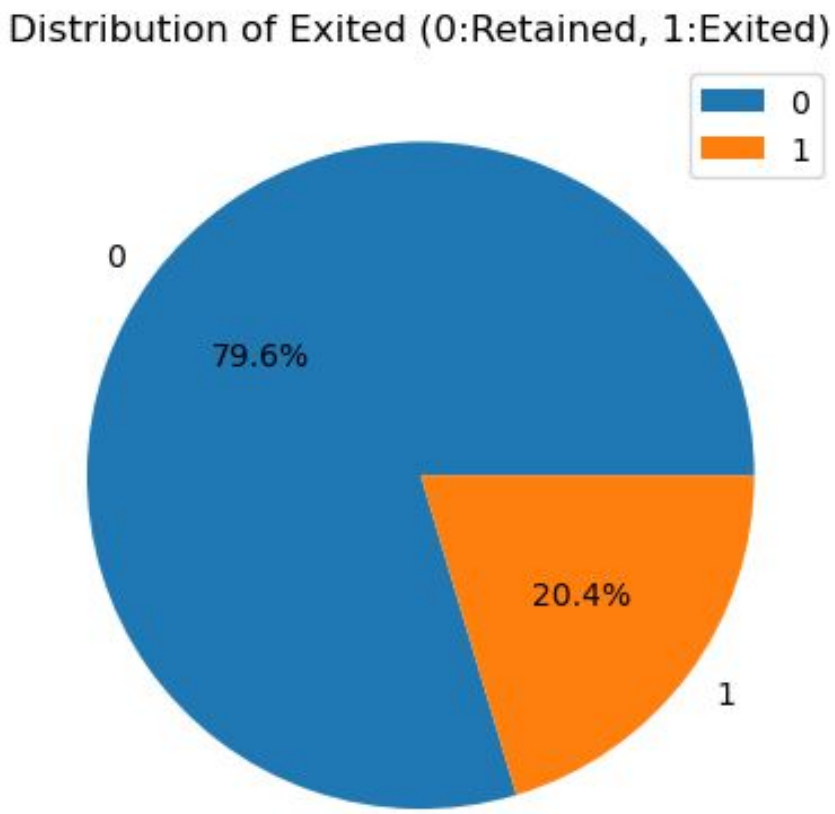
	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
0	1	15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	101348.88	1
1	2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
2	3	15619304	Onio	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
3	4	15701354	Boni	699	France	Female	39	1	0.00	2	0	0	93826.63	0
4	5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
9995	9996	15606229	Obijiaku	771	France	Male	39	5	0.00	2	1	0	96270.64	0
9996	9997	15569892	Johnstone	516	France	Male	35	10	57369.61	1	1	1	101699.77	0
9997	9998	15584532	Liu	709	France	Female	36	7	0.00	1	0	1	42085.58	1
9998	9999	15682355	Sabbatini	772	Germany	Male	42	3	75075.31	2	1	0	92888.52	1
9999	10000	15628319	Walker	792	France	Female	28	4	130142.79	1	1	0	38190.78	0

10000 rows × 14 columns

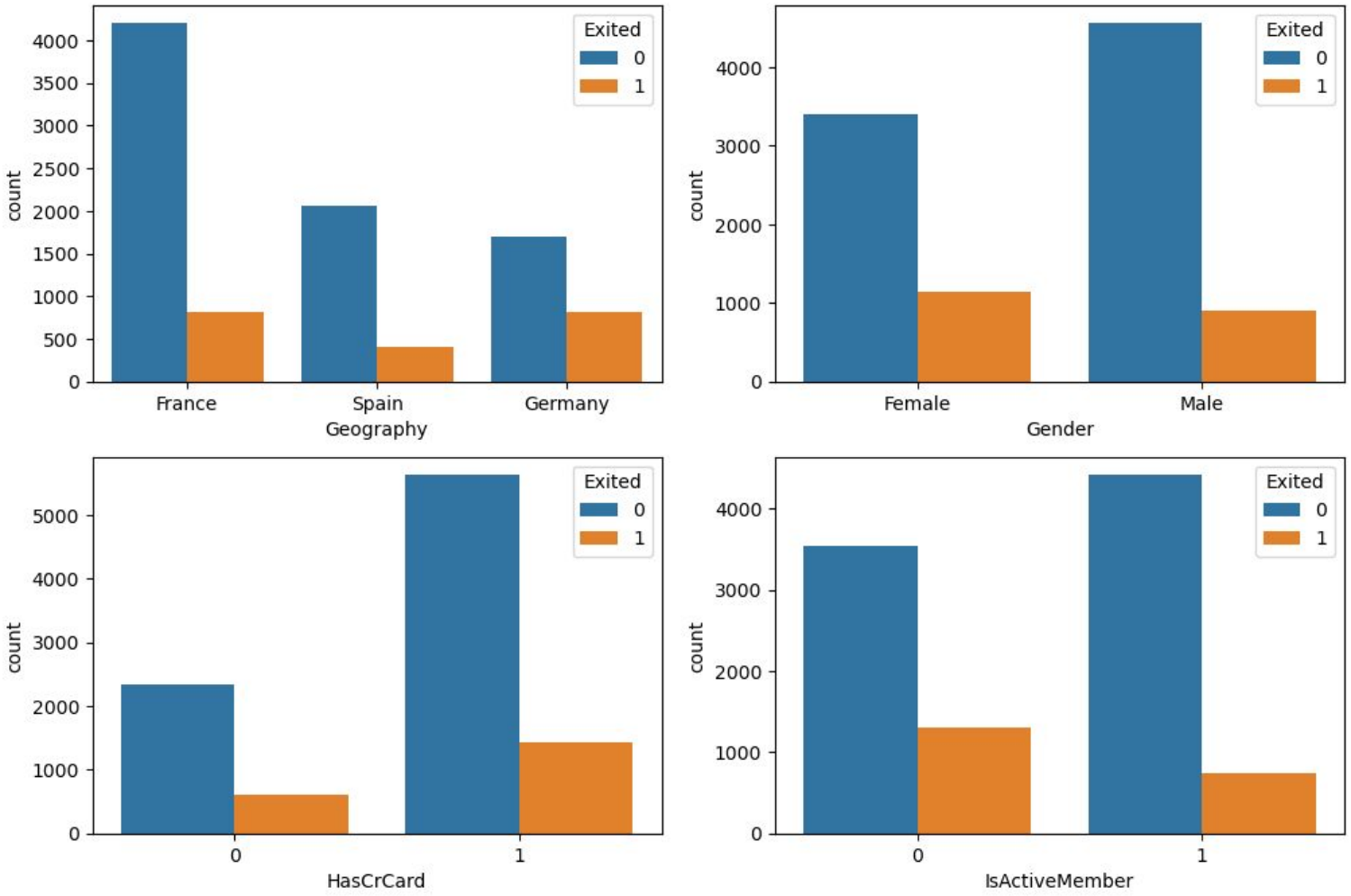
https://www.kaggle.com/datasets/adammaus/predicting-churn-for-bank-customers

Phân tích dữ liệu

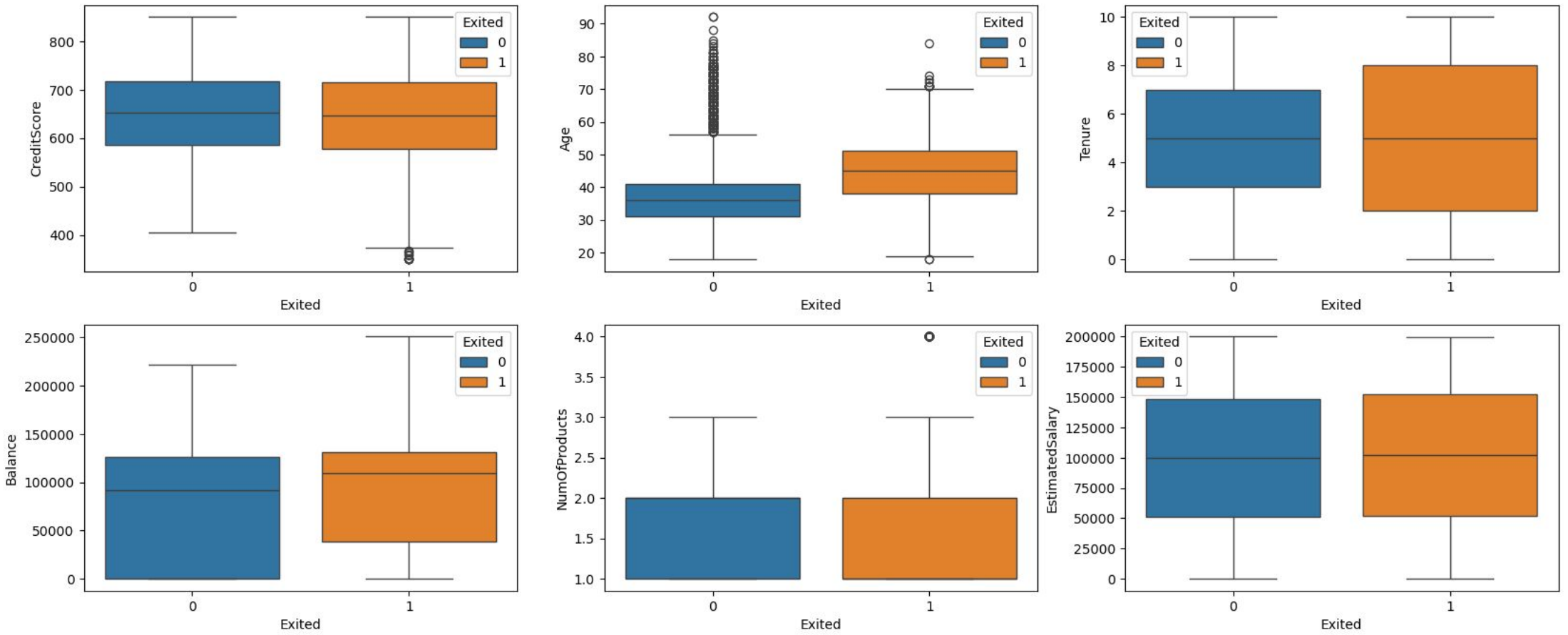
Nhận xét: Tập dữ liệu mất cân bằng (Exited ~ 20,4%)



Phân tích dữ liệu



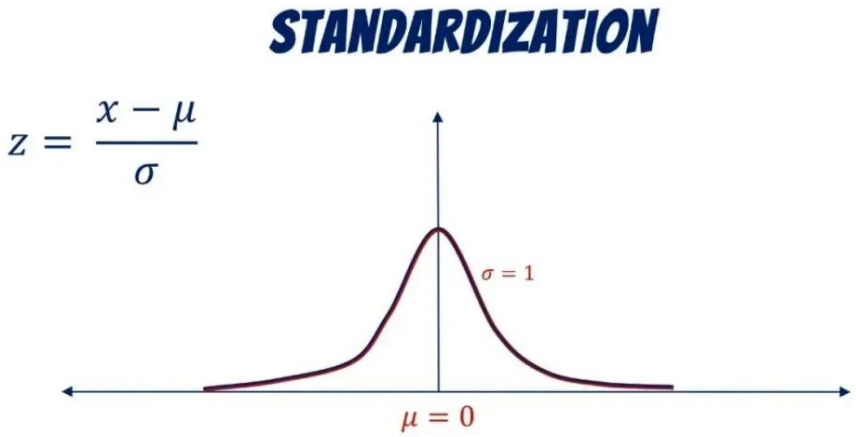
Phân tích dữ liệu



Tiền xử lý dữ liệu

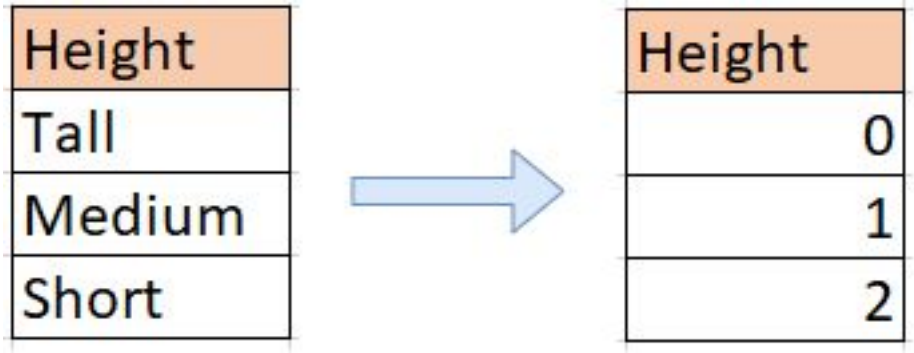
- **Feature Scaling:**

Standardization: là một kỹ thuật điều chỉnh trong đó các giá trị trung bình bằng 0 với độ lệch chuẩn bằng một.



- **Xử lý dữ liệu categorical:**

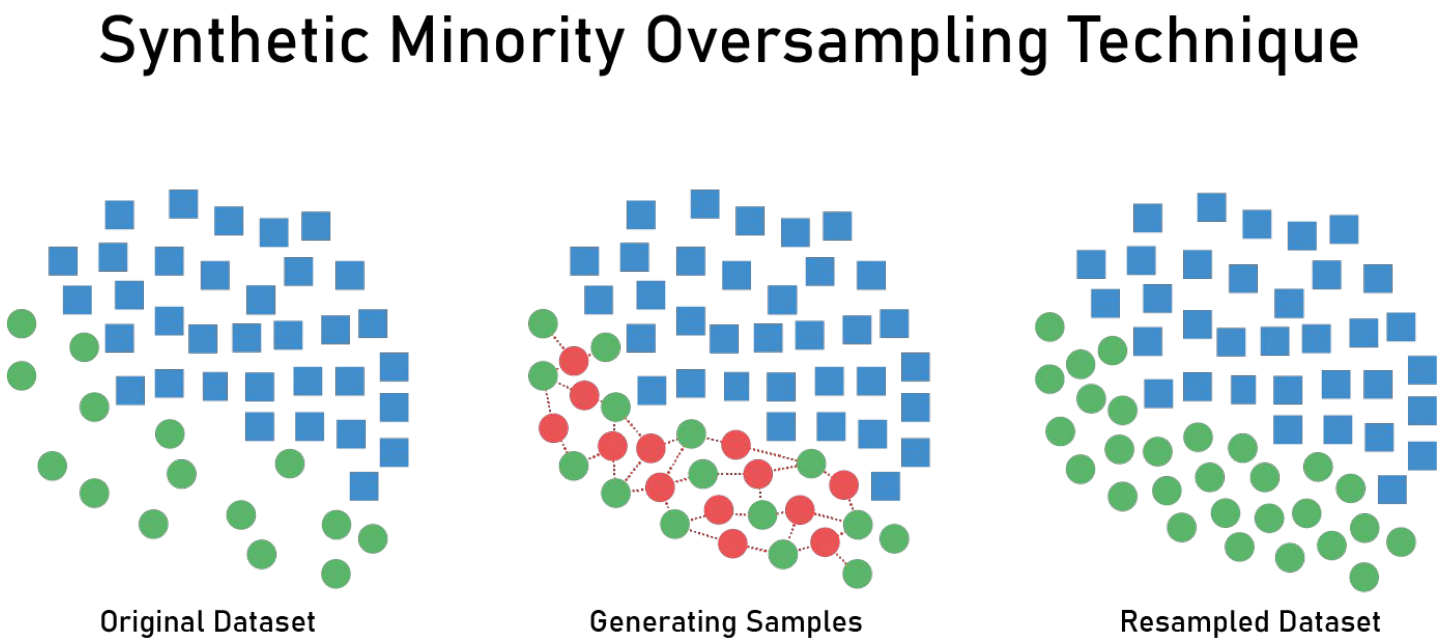
Label encoder: là kỹ thuật tiền xử lý dữ liệu trong học máy dùng để chuyển đổi các biến phân loại (categorical) thành dạng số nguyên (numerical)



Xử lý dữ liệu mất cân bằng

Smote: (Kỹ thuật lấy mẫu quá mức thiểu số tổng hợp – xử lý mất cân bằng dữ liệu).

Phương pháp này tạo ra các ví dụ tổng hợp của lớp thiểu số để cân bằng tập dữ liệu.



Cài đặt mô hình

Trước khi xử lý mất cân bằng dữ liệu

```
lr = LogisticRegression()
lr.fit(X_train, y_train)
y_pred = lr.predict(X_test)
print('Accuracy:', accuracy_score(y_test, y_pred))
print('Precision:', precision_score(y_test, y_pred))
print('Recall:', recall_score(y_test, y_pred))
print('F1 Score:', f1_score(y_test, y_pred))
print('Classification Report:\n', classification_report(y_test, y_pred))
```

✓ 0.0s

```
Accuracy: 0.815
Precision: 0.5966386554621849
Recall: 0.1806615776081425
F1 Score: 0.27734375
Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.97      0.89      1607
           1       0.60      0.18      0.28       393

    accuracy                           0.81      2000
   macro avg       0.71      0.58      0.59      2000
weighted avg       0.78      0.81      0.77      2000
```

Sau khi xử lý mất cân bằng dữ liệu

```
lr = LogisticRegression()
lr.fit(X_res, y_res)
y_pred = lr.predict(X_test)
print('Accuracy:', accuracy_score(y_test, y_pred))
print('Precision:', precision_score(y_test, y_pred))
print('Recall:', recall_score(y_test, y_pred))
print('F1 Score:', f1_score(y_test, y_pred))
print('Classification Report:\n', classification_report(y_test, y_pred))
```

✓ 0.0s

```
Accuracy: 0.714
Precision: 0.3712230215827338
Recall: 0.6564885496183206
F1 Score: 0.4742647058823529
Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.73      0.80      1607
           1       0.37      0.66      0.47       393

    accuracy                           0.71      2000
   macro avg       0.63      0.69      0.64      2000
weighted avg       0.79      0.71      0.74      2000
```

Cài đặt mô hình

Suport Vector Machine

```
svc = SVC(probability=True)
svc.fit(X_res, y_res)
y_pred = svc.predict(X_test)
print('Accuracy:', accuracy_score(y_test, y_pred))
print('Precision:', precision_score(y_test, y_pred))
print('Recall:', recall_score(y_test, y_pred))
print('F1 Score:', f1_score(y_test, y_pred))
print('Classification Report:\n', classification_report(y_test, y_pred))
```

[35] ✓ 28.9s

```
Accuracy: 0.7945
Precision: 0.48509933774834435
Recall: 0.7455470737913485
F1 Score: 0.5877632898696088
Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.81      0.86      1607
           1       0.49      0.75      0.59       393

    accuracy                           0.79      2000
   macro avg       0.71      0.78      0.73      2000
weighted avg       0.84      0.79      0.81      2000
```

Random Forest

```
rf = RandomForestClassifier()
rf.fit(X_res, y_res)
y_pred = rf.predict(X_test)
print('Accuracy:', accuracy_score(y_test, y_pred))
print('Precision:', precision_score(y_test, y_pred))
print('Recall:', recall_score(y_test, y_pred))
print('F1 Score:', f1_score(y_test, y_pred))
print('Classification Report:\n', classification_report(y_test, y_pred))
```

[36] ✓ 2.1s

```
Accuracy: 0.844
Precision: 0.6025316455696202
Recall: 0.6055979643765903
F1 Score: 0.6040609137055838
Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.90      0.90      1607
           1       0.60      0.61      0.60       393

    accuracy                           0.84      2000
   macro avg       0.75      0.75      0.75      2000
weighted avg       0.84      0.84      0.84      2000
```

Đánh giá mô hình

Model	Accuracy	Precision	Recall	F1
Logistic Regression	0,714	0,37	0,66	0,47
SVM	0,795	0,49	0,75	0,59
Random Forest	0,844	0,6	0,61	0,6

Contents

Thank You !

Question?