Mining Massive Datasets

Người thực hiện
Đặng Quý Anh
Trần Thanh Tùng
Vũ Thị Thùy Dung

Mục lục

Mục lục

Thành viên nhóm và phân chia công việc

- 19000244 - Đặng Quý Anh - Tiền xử lý và huấn luyện mô hình Support Vector Machines
- 19000304 - Trần Thanh Tùng - Tìm hiểu các metric đánh giá và huấn luyện mô hình Decision Tree
- 19000403 - Vũ Thị Thùy Dung - Tìm hiểu về pyspark và huấn luyện mô hình K-means

Mục lục

Giảm kích thước dữ liệu

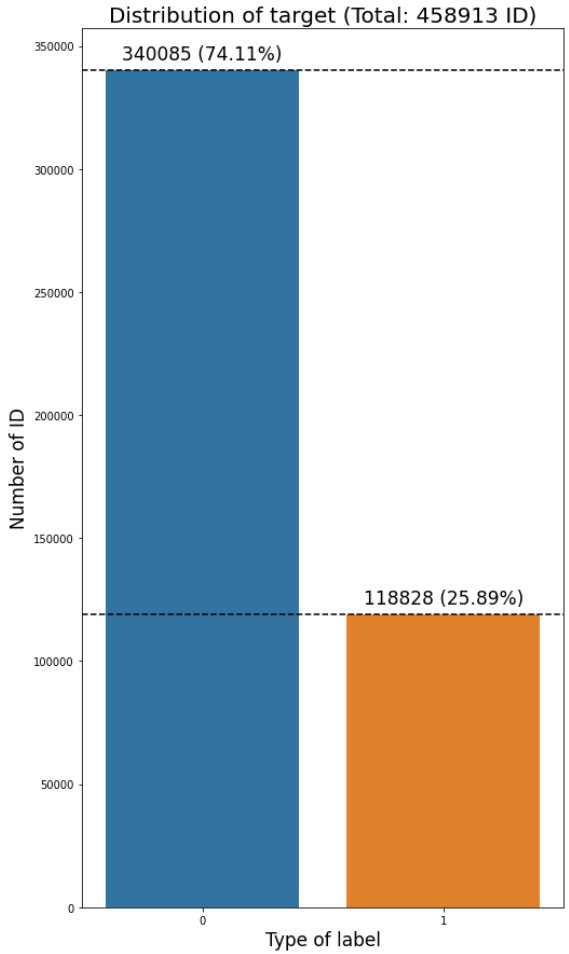
Hướng tiếp cận

- Thay đổi định dạng tập tin: csv $\Rightarrow$ parquet
- Thay đổi kiểu của trường dữ liệu

Tên file	Trước khi xử lý	Sau khi xử lý
train_data	16.39 GB	1.53 GB
test_data	33.82 GB	3.07 GB

Khai phá dữ liệu

- Có 190 biến thể hiện đăc trưng, 1 biến mục tiêu
- Trong 190 biến có 11 biến categorical, 1 biến thể hiện customer_ID, 1 biến thể hiện ngày, còn lại là biến liên tục (numeric)
- Biến mục tiêu chứa 2 giá trị là 0 và 1



Xử ký giá trị bị thiếu

- 61/190 trường dữ liệu là có giá trị null
- Các trường giá trị null chiếm trên 60% sẽ bị loại bỏ
- Với các trường còn lại: biến categorical (lấy theo mode), biến numeric (lấy theo mean) để điền vào giá trị null

Xử lý giá trị bị thiếu

	Quantity	Percentage	Type
D_88	5525447	99.89	float32
D_110	5500117	99.43	float32
B_39	5497819	99.39	float32
D_73	5475595	98.99	float32
B_42	5459973	98.71	float32
D_134	5336752	96.48	float32
B_29	5150035	93.10	float32
D_132	4988874	90.19	float32
D_76	4908954	88.75	float32
D_42	4740137	85.69	float32
D_142	4587043	82.93	float32
D_53	4084585	73.84	float32

Hình 2: Chi tiết các trường dữ liệu có giá trị null chiếm hơn 60%

Xử lý giá trị bị trùng lặp

- Có 5531451 giao dịch của 458913 khác hàng theo ngày
- Cần nhóm các giao dịch này theo ID để tiếp tục giảm kích thước
  - Lấy tổng trên các trường liên tục
  - Lấy max trên các trường rời rạc

Mục lục

Một số metric để đánh giá mô hình

Sử dụng ma trận nhầm lẫn (confusion matrix)

	Actual Positive	**Actual Negative**
Predicted Positive	True Positives (TP)	False Positives (FP)
Predicted Negative	False Negatives (FN)	True Negatives (TN)

Trong đó,

- True Positives (TP): Đây là các trường hợp mà mô hình dự đoán lớp poitive và lớp thực tế cũng là positive.
- True Negatives (TN): Đây là các trường hợp mà mô hình dự đoán lớp negative và lớp thực tế cũng là negative.
- False Positives (FP): Đây là các trường hợp mà mô hình dự đoán lớp positive nhưng lớp thực tế là negative.
- False Negatives (FN): Đây là các trường hợp mà mô hình dự đoán lớp negative nhưng lớp thực tế là poitive.

Một số metric để đánh giá mô hình

Từ ma trận nhầm lẫn trên

- **Accuracy**:

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

- **Precision**:

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

- **Recall**:

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1**:

$$\text{F1 Score} = \frac{2 * (Precision * Recall)}{Precision + Recall}$$

Mục lục

Song song hóa với pyspark

RDD (Resilient Distributed Datasets) là cấu trúc dữ liệu cơ bản của Spark và là nguồn dữ liệu phân tán chính trong ứng dụng Spark, với một số tính chất sau

- **Immutable**: Sau khi RDD được tạo, nó không thể sửa đổi được.
- **Distributed**: RDD được phân phối trên một cụm máy, cho phép chúng được xử lý song song.
- **Fault-tolerant**: RDD có thể được tính toán lại nếu một phần lưu trữ nút của RDD bị lỗi hoặc nếu tập dữ liệu cần được tính toán lại vì bất kỳ lý do gì.
- **Parallel**: RDD có thể được xử lý song song, điều này làm cho chúng hiệu quả để xử lý dữ liệu lớn.

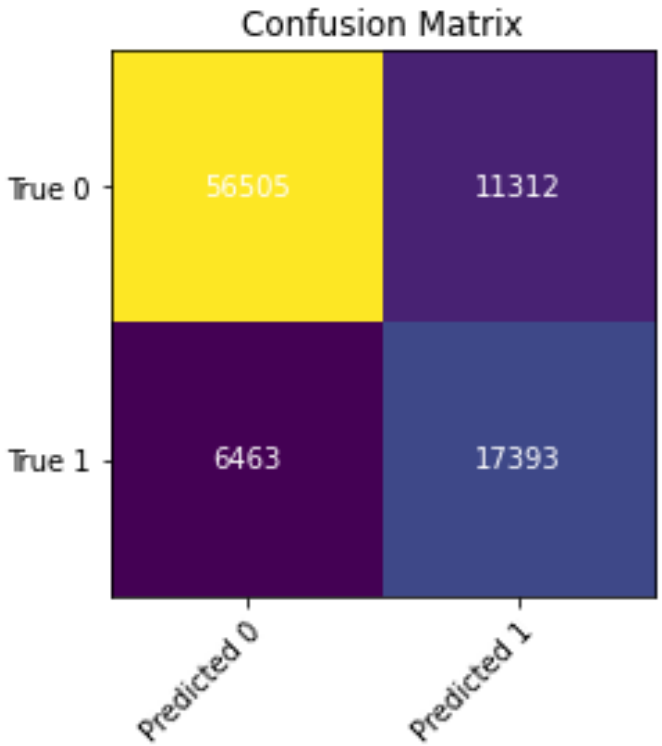
Song song hóa với pyspark

Các bước để triên khai mô hình

- Dữ liệu từ tập tin parquet đã được xử lý, đọc qua thư viện pyspark
- Sử dụng RDD trên pyspark để song song hóa
- Chia dữ liệu đầu vào thành 2 phần: 80% để traning, 20% để validation
- Đưa vào các mô hình Kmeans, Decision Tree và SVMs để huấn luyện

Mục lục

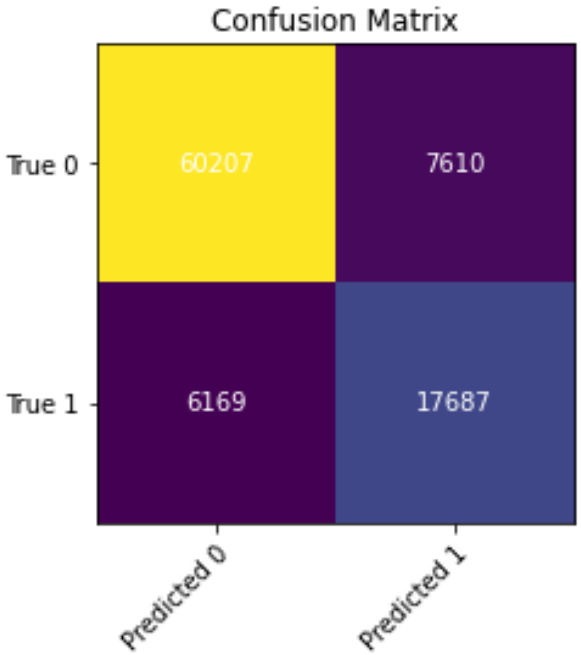
Mô hình K-means



	precision	recall	f1-score	support
0.0	0.90	0.83	0.86	67817
1.0	0.61	0.73	0.66	23856
accuracy			0.81	91673
macro avg	0.75	0.78	0.76	91673
weighted avg	0.82	0.81	0.81	91673

Hình 3: Confusion matrix và một số metric đánh giá

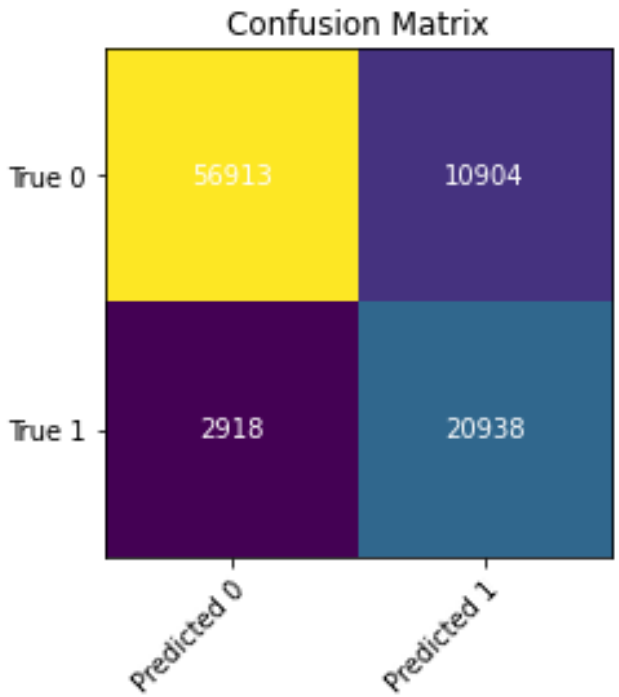
Mô hình Decision Tree



	precision	recall	f1-score	support
0.0	0.91	0.89	0.90	67817
1.0	0.70	0.74	0.72	23856
accuracy			0.85	91673
macro avg	0.80	0.81	0.81	91673
weighted avg	0.85	0.85	0.85	91673

Hình 4: Confusion matrix và một số metric đánh giá

Mô hình SVMs



	precision	recall	f1-score	support
0.0	0.95	0.84	0.89	67817
1.0	0.66	0.88	0.75	23856
accuracy			0.85	91673
macro avg	0.80	0.86	0.82	91673
weighted avg	0.87	0.85	0.86	91673

Hình 5: Confusion matrix và một số metric đánh giá

Mục lục

Tài liệu tham khảo

[1] Aurélien Géron, "Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow", O'reilly, 2019.
[2] Marc Peter Deisenroth, A. Aldo Faisal, Cheng Soon Ong, "Mathematics for Machine Learning", Cambridge University Press, 2020.
[3] "K-Means Clustering in Python: A Practical Guide" - https://realpython.com/k-means-clustering-python/
[4] "Clustering" - https://scikit-learn.org/stable/module/clustering.htmlk-means
[5] "Blog: Machine Learning cơ bản" - https://machinelearningcoban.com

Tài liệu tham khảo

[6] Nagesh Singh Chauhan, "Decision Tree Algorithm, Explained" - https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html
[7] "Decision tree learning" - https://en.wikipedia.org/wiki/Decision_tree_learning
[8] "ID3 algorithm" - https://en.wikipedia.org/wiki/ID3_algorithm
[9] "Decision Trees" - https://scikit-learn.org/stable/modules/tree.html
[10] "Support Vector Machines" - https://en.wikipedia.org/wiki/Support_vector_machine
[11] "Support Vector Machines" - https://scikit-learn.org/stable/modules/svm.html
[12] Dataset - https://www.kaggle.com/competitions/amex-default-prediction/data
[13] MapReduce - https://en.wikipedia.org/wiki/MapReduce

Tài liệu tham khảo

[11] "Support Vector Machines" - https://scikit-learn.org/stable/modules/svm.html
[12] Dataset - https://www.kaggle.com/competitions/amex-default-prediction/data
[13] MapReduce - https://en.wikipedia.org/wiki/MapReduce