**Phát hiện gian lận khi tự checkout ở cửa hàng bán lẻ**

Số lượng các trạm tự checkout đang tăng lên. Điều này bao gồm (1) tự checkout tại chỗ (khách hàng mang giỏ hàng của họ đến một trạm quét và thanh toán cho sản phẩm của họ) và (2) tự checkout bán cố định (khách hàng quét sản phẩm trực tiếp và chỉ thanh toán tại quầy). Khách hàng sử dụng điện thoại thông minh của riêng họ để quét hoặc cửa hàng cung cấp máy quét di động.

Quy trình tự động này giúp tránh phải xếp hàng dài và tăng tốc quá trình thanh toán cho khách hàng cá nhân. Nhưng làm thế nào các nhà bán lẻ có thể ngăn việc niềm tin mà họ đã đặt vào khách hàng bị lạm dụng? Làm thế nào họ có thể quyết định đơn hàng nào cần kiểm tra phát hiện những khách hàng gian lận mà không làm phiền khách hàng trong sạch?

**Tình huống**

Một nhà bán lẻ thực phẩm đã giới thiệu hệ thống tự quét (scan) cho phép khách hàng quét các mặt hàng của họ bằng máy quét di động cầm tay trong khi mua sắm.

Hình thức thanh toán này khiến các nhà bán lẻ gặp rủi ro rằng một số lượng khách hàng nhất định sẽ lợi dụng quyền tự do này để thực hiện hành vi gian lận bằng cách không quét tất cả các mặt hàng trong giỏ hàng của họ.

Nghiên cứu thực nghiệm được thực hiện bởi các nhà cung cấp đã chỉ ra rằng sự khác biệt được tìm thấy trong khoảng 5% của tất cả các giao dịch tự quét. Nghiên cứu không phân biệt giữa mục đích gian lận thực tế của khách hàng, lỗi vô ý hoặc sự cố kỹ thuật với máy quét.

Để giảm thiểu thiệt hại, nhà bán lẻ thực phẩm hy vọng sẽ xác định được các trường hợp gian lận bằng cách sử dụng các biện pháp kiểm tra theo dõi có mục tiêu. Thách thức ở đây là giữ cho số lần kiểm tra càng thấp càng tốt để tránh tăng thêm chi phí không cần thiết cũng như tránh gây phiền hà cho những khách hàng vô tội do bị buộc tội sai. Tuy nhiên, đồng thời, mục tiêu là xác định càng nhiều lần quét sai càng tốt.

Mục tiêu của các nhóm là tạo ra một mô hình để phân loại các lần quét là gian lận hoặc không gian lận. Việc phân loại không tính đến việc gian lận được thực hiện một cách cố ý hay vô ý.

Để tạo mô hình này, các nhóm nhận thông tin về các lần quét và phân loại của chúng dùng tập huấn luyện.

**Dữ liệu**

Dữ liệu được ẩn danh và ở dạng có cấu trúc (csv) được cung cấp cho nhiệm vụ này. Dưới đây là một số điểm cần lưu ý về các file dữ liệu:

1. Mỗi đối tượng dữ liệu nằm trên một dòng, kết thúc bằng "LF" ("xuống dòng”).
2. Dòng đầu tiên là header, chứa tên của các cột tương ứng.
3. Ở mỗi dòng, mỗi giá trị cách xa nhau bằng dấu "|".
4. Các "." được sử dụng làm dấu phân cách thập phân. Giá trị không được làm tròn.
5. Không có ký tự thoát, không sử dụng dấu ngoặc kép.
6. ASCII là bộ ký tự được sử dụng.

Danh sách tất cả các tên cột (thuộc tính) cũng như các mô tả và miền giá trị của chúng như sau:

|  |  |  |
| --- | --- | --- |
| Tên cột (thuộc tính) | Mô tả | Miền giá trị |
| trustLevel | Mức độ tin cậy của khách hàng  6: mức độ cao nhất | {1,2,3,4,5,6} |
| totalScanTimeInSeconds | Tổng thời gian theo giây giữa sản phẩm đầu với sản phẩm cuối được quét (scan) | Số nguyên dương |
| grandTotal | Tổng số sản phẩm đã được quét | Số thập phân dương với tối đa hai chữ số thập phân |
| lineItemVoids | Số lần quét vô hiệu (không hợp lệ) | Số nguyên dương |
| scansWithoutRegistration | Số lần kích hoạt máy quét mà không thực sự quét bất kỳ thứ gì | Số nguyên dương hoặc 0 |
| quantityModification | Số lượng đã sửa đổi cho một trong các sản phẩm được quét | Số nguyên dương hoặc 0 |
| scannedLineItemsPerSecond | Số lượng sản phẩm được quét trung bình mỗi giây | Số thập phân dương |
| valuePerSecond | Giá trị trung bình của các sản phẩm được quét mỗi giây | Số thập phân dương |
| lineItemVoidsPerPosition | Số lượng hàng bị hỏng trung bình trên tổng số tất cả các sản phẩm đã quét và chưa hủy | Số thập phân dương |
| fraud | Phân loại là gian lận (1) hoặc không gian lận (0) | {0,1} |

Tập “train.csv” được sử dụng để huấn huyện mô hình. Tập “test.csv” được sử dụng cho dự đoán.

Một dòng dữ liệu trong tập “train.csv” hoặc “test.csv” chứa thông tin về một quy trình quét cho một khách hàng, bao gồm số lượng mặt hàng được quét mỗi giây và tổng thời gian ở cửa hàng.

Tập “train.csv” cũng chứa một cột fraud có phân loại là gian lận và không gian lận.

**Nhiệm vụ**

Sử dụng dữ liệu được cho để tạo mô hình nhằm phân loại các lần quét là gian lận hay không gian lận.

Một tập chứa thông tin sau sẽ được sử dụng để gửi kết quả:

|  |  |  |
| --- | --- | --- |
| Tên cột | Mô tả | Miền giá trị |
| fraud | Phân loại là gian lận (1) hoặc không gian lận (0) | {0, 1} |

Không có thuộc tính khóa. Số lượng và thứ tự của các dự đoán phải tương ứng với các thứ tự trong tập “test.csv”. Các giá trị có thể có cho cột " fraud" là các giá trị nguyên 0 hoặc 1. File kết quả có thể trông như sau:

fraud

1

0

0

…

File kết quả phải tuân thủ các thông số kỹ thuật được mô tả trong phần dữ liệu và phải được tải lên dưới dạng có cấu trúc (csv) lên trang web UTEXLMS. Tên của file kết quả bao gồm tên nhóm: “**tennhom.csv**“ Các trưởng nhóm đăng ký tên nhóm với GV.

**Đánh giá**

Kết quả của các nhóm sẽ được đánh giá và so sánh dựa trên ma trận chi phí (cost matrix) sau:

|  |  |  |  |
| --- | --- | --- | --- |
| Prediction | Actual value | | |
|  | 0 (no fraud) | 1 (fraud) |
| 0 (no fraud) | 0 | -5 |
| 1 (fraud) | -25 | 5 |

Nhà bán lẻ thực phẩm nhận được lợi nhuận là 5 cho mỗi lần gian lận được xác định chính xác. Tuy nhiên, đối với mọi trường hợp gian lận mà không bị phanh phui, anh ta sẽ mất 5.

Một người bán hàng bị buộc tội gian lận, có thể không quay lại cửa hàng này, điều này được biểu thị bằng khoản lỗ 25 cho nhà bán lẻ.

Một khách hàng trung thực được xác định một cách chính xác có nghĩa là nhà bán lẻ không có lãi hay lỗ. Tổng chi phí hoặc lợi nhuận của tất cả các lần quét là giá trị của giải pháp đã gửi.

Đội chiến thắng là đội có giải pháp thu được lợi nhuận bằng tiền cao nhất.