**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**ĐỒ ÁN CUỐI KÌ**

**NHẬP MÔN KĨ THUẬT DỮ LIỆU**

*Người thực hiện*: **VÕ TRỌNG NHƠN – 22658441**

**TRẦN XUÂN DIỆN – 22650601**

**TRẦN PHÚ THỌ – 22653431**

**ĐỖ TẤN ĐẠT – 22648601**

**NGUYỄN ĐĂNG TUẤN HUY – 22658341**

Lớp **: DHKHDL18A – 420301541001**

Khoá  **: 18**

*Người hướng dẫn*: **TS NGUYỄN CHÍ KIÊN**

**ThS. TRƯƠNG VĨNH LINH**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024**

**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**ĐỒ ÁN CUỐI KÌ**

**NHẬP MÔN KĨ THUẬT DỮ LIỆU**

*Người thực hiện*: **VÕ TRỌNG NHƠN – 22658441**

**TRẦN XUÂN DIỆN – 22650601**

**TRẦN PHÚ THỌ – 22653431**

**ĐỖ TẤN ĐẠT – 22648601**

**NGUYỄN ĐĂNG TUẤN HUY – 22658341**

Lớp **: DHKHDL18A – 420301541001**

Khoá  **: 18**

*Người hướng dẫn*: **TS NGUYỄN CHÍ KIÊN**

**ThS. TRƯƠNG VĨNH LINH**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024**

**LỜI CẢM ƠN**

Trong báo cáo cuối kỳ này, chúng em xin phép được bày tỏ lòng biết ơn sâu sắc đến những người đã hỗ trợ và tạo điều kiện thuận lợi cho sự phát triển của chúng em trong môn Kĩ Thuật Dữ Liệu.

Trước hết, chúng em xin gửi lời cảm ơn chân thành đến thầy Kiên và thầy Linh. Thầy đã luôn hướng dẫn, chia sẻ kiến thức và kinh nghiệm quý báu, giúp chúng em không chỉ hiểu sâu hơn về lĩnh vực dữ liệu mà còn phát triển kỹ năng thực hành và tư duy phân tích. Sự tận tâm và nhiệt huyết của thầy đã là nguồn động lực lớn lao, giúp chúng em vượt qua những khó khăn và thử thách trong suốt quá trình thực hiện đồ án.

Chúng em cũng xin bày tỏ lòng biết ơn đến chính bản thân mình, vì đã không ngừng cố gắng, làm việc chăm chỉ và kiên trì để hoàn thành đồ án này. Tinh thần làm việc nhóm, sự cống hiến và tinh thần trách nhiệm của mỗi thành viên trong nhóm đã góp phần quan trọng vào sự thành công của dự án.

Cuối cùng, chúng em xin gửi lời chúc sức khỏe và lời chúc tốt đẹp nhất đến các thầy. Hy vọng các thầy sẽ tiếp tục dẫn dắt và hỗ trợ thế hệ sinh viên tiếp theo, giúp họ đạt được nhiều thành tựu hơn nữa trong học tập và nghiên cứu. Chúng em cũng chúc cho các nhóm dự án cuối kỳ khác sẽ thành công rực rỡ và đạt được những kết quả xuất sắc.

Xin chân thành cảm ơn!

**ĐỒ ÁN ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP HỒ CHÍ MINH**

Tôi xin cam đoan đây là sản phẩm đồ án của chúng tôi và được sự hướng dẫn của TS. Nguyễn Chí Kiên và ThS. Trương Vĩnh Linh. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây.

**Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình.** Trường đại học Công nghiệp TP Hồ Chí Minh không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày tháng năm*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

*Võ Trọng Nhơn*

*Trần Xuân Diện*

*Trần Phú Thọ*

*Đỗ Tấn Đạt*

*Nguyễn Đăng Tuấn Huy*

**PHẦN ĐÁNH GIÁ CỦA GIẢNG VIÊN**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

**MỤC LỤC**

I. GIỚI THIỆU

1.1 Mục tiêu của báo cáo

1.2. Tầm quan trọng của việc dự đoán điểm khách hàng tiềm năng

II. DỮ LIỆU VÀ TIỀN XỬ LÝ

2.1. Mô tả dữ liệu

2.2. Các bước tiền xử lý dữ liệu

2.3. Chia dữ liệu thành tập huấn luyện và tập kiểm tra

III. PHƯƠNG PHÁP DỰ ĐOÁN

3.1. Các thuật toán sử dụng

3.2. Hàm predict\_proba

3.3. Quy đổi xác suất thành điểm

IV. ỨNG DỤNG THỰC TẾ

4.1. Xác định các tương tác quan trọng

4.2. Chấm điểm khách hàng tiềm năng

4.3. Tối ưu hóa chiến lược tái mục tiêu

V. PIPELINE

5.1. Các bước và tính năng của pipeline

5.2. Đánh giá pipeline

VI. DEMO QUẢN LÝ KHÁCH HÀNG

6.1. Hệ thống đăng ký người dùng

6.2. Bảng điều khiển quản lý

6.3. Tải dữ liệu mô phỏng

6.4. Mô phỏng tình huống

6.5. Database

6.6. Xử lý lỗi

VI. KẾT LUẬN

7.1. Tóm tắt kết quả

7.2. Hướng phát triển trong tương lai

**CHƯƠNG I**

**GIỚI THIỆU**

**1.1. Mục tiêu của báo cáo**

Mục tiêu của báo cáo này là trình bày phương pháp và kết quả của việc dự đoán điểm của khách hàng tiềm năng. Việc dự đoán điểm khách hàng không chỉ giúp doanh nghiệp phân loại và quản lý khách hàng một cách hiệu quả mà còn cho phép tối ưu hóa các chiến lược marketing và bán hàng. Bằng cách sử dụng các kỹ thuật học máy và phân tích dữ liệu tiên tiến, chúng tôi nhằm mục đích xây dựng một mô hình dự đoán có khả năng xác định chính xác những khách hàng có khả năng chuyển đổi cao nhất. Kết quả của nghiên cứu này sẽ cung cấp cơ sở dữ liệu để doanh nghiệp có thể đưa ra quyết định chiến lược, nâng cao hiệu quả kinh doanh và cải thiện mối quan hệ với khách hàng.

Ngoài ra, chúng tôi cũng đã tạo ra một trang web demo quản lý khách hàng với các chức năng cơ bản như nhập dữ liệu và trực quan hóa dữ liệu của một số khách hàng tiềm năng. Trang web này cho phép người dùng dễ dàng quản lý thông tin khách hàng và theo dõi kết quả dự đoán, từ đó hỗ trợ trong việc đưa ra các quyết định kinh doanh một cách nhanh chóng và chính xác.

**1.2. Tầm quan trọng của việc dự đoán điểm khách hàng tiềm năng**

Trong bối cảnh cạnh tranh ngày càng gia tăng, việc xác định chính xác khách hàng tiềm năng có thể mang lại nhiều lợi ích quan trọng cho doanh nghiệp. Dự đoán điểm khách hàng tiềm năng giúp doanh nghiệp tập trung nguồn lực vào những khách hàng có giá trị cao, từ đó nâng cao tỷ lệ chuyển đổi và giảm chi phí marketing.

**1. Tối ưu hóa chi phí marketing:**

- Thay vì đầu tư vào các chiến dịch marketing rộng khắp và tốn kém, doanh nghiệp có thể tập trung vào các nhóm khách hàng có tiềm năng cao nhất. Điều này giúp giảm chi phí marketing và tối đa hóa hiệu quả đầu tư (ROI).

**2. Nâng cao tỷ lệ chuyển đổi:**

- Bằng cách nhận diện đúng khách hàng tiềm năng, doanh nghiệp có thể thiết kế các chiến dịch marketing và bán hàng phù hợp, tăng khả năng chuyển đổi từ khách hàng tiềm năng thành khách hàng thực tế.

**3. Cải thiện trải nghiệm khách hàng:**

- Hiểu rõ hơn về nhu cầu và hành vi của khách hàng tiềm năng giúp doanh nghiệp cung cấp các sản phẩm và dịch vụ phù hợp hơn, từ đó nâng cao mức độ hài lòng và trung thành của khách hàng.

**4. Tăng cường hiệu quả kinh doanh:**

- Việc dự đoán chính xác khách hàng tiềm năng cho phép doanh nghiệp dự báo doanh thu chính xác hơn, quản lý hàng tồn kho hiệu quả và lập kế hoạch sản xuất linh hoạt.

**5. Phân bổ nguồn lực hợp lý:**

- Doanh nghiệp có thể phân bổ nguồn lực (nhân lực, tài chính, thời gian) một cách hiệu quả hơn, tập trung vào các hoạt động có giá trị gia tăng cao và bỏ qua những khách hàng có khả năng thấp hơn.

Trong thời đại số hóa, việc áp dụng các kỹ thuật phân tích dữ liệu và học máy để dự đoán điểm khách hàng tiềm năng không chỉ là một lợi thế cạnh tranh mà còn là một yếu tố quyết định sự thành công của doanh nghiệp. Nghiên cứu này không chỉ có ý nghĩa trong việc cải thiện hiệu quả kinh doanh mà còn đóng góp vào sự phát triển của khoa học dữ liệu ứng dụng trong lĩnh vực marketing và quản lý khách hàng.

**CHƯƠNG II**

**DỮ LIỆU VÀ TIỀN XỬ LÝ**

**2.1. Mô tả dữ liệu**

Dữ liệu sử dụng trong nghiên cứu này bao gồm các thông tin cơ bản về khách hàng dựa trên các hoạt động trên trang web mà khách hàng tương tác. Các biến số được thu thập nhằm cung cấp cái nhìn tổng quan về hành vi và đặc điểm của khách hàng. Cụ thể, dữ liệu chứa các thông tin như số lần tương tác vào biểu tượng giỏ hàng, số lần thêm vào danh sách giỏ hàng, số lần xem chi tiết giỏ hàng, các thao tác sắp xếp, lựa chọn hình ảnh, nhấp vào trang tài khoản, xem chi tiết giao hàng, xem chi tiết trả hàng, đăng nhập, xem trang thanh toán, xem bảng kích thước, và nhiều hành động khác trên trang web. Ngoài ra, dữ liệu còn bao gồm các biến liên quan đến loại thiết bị sử dụng (máy tính, máy tính bảng, thiết bị di động), người dùng quay lại, và vị trí địa lý của người dùng

**2.2. Các bước tiền xử lý dữ liệu**

**1. Loại bỏ các biến không cần thiết:**

- Chúng tôi loại bỏ biến *'ordered'* vì đây là biến mục tiêu trong mô hình dự đoán của chúng tôi.

- Biến *'UserID'* cũng bị loại bỏ vì không ảnh hưởng trực tiếp đến khả năng đặt hàng của khách hàng.

- Biến *'device\_mobile'* được loại bỏ do có mối tương quan nghịch với các đơn đặt hàng, điều này có thể gây nhiễu cho mô hình dự đoán.

**2. Xác định biến dự đoán và biến mục tiêu:**

- Các biến dự đoán được sử dụng để xây dựng mô hình bao gồm: *'checked\_delivery\_pdp'*, *'basket\_icon\_click'*, *'sign\_in'*, *'saw\_checkout'*, và nhiều biến khác trừ các biến đã loại bỏ ở bước trên.

- Biến mục tiêu là *'ordered'*, đại diện cho hành động đặt hàng của khách hàng.

**3. Xử lý các giá trị thiếu và dữ liệu ngoại lai:**

- Mặc dù dữ liệu từ Kaggle đã được làm sạch, chúng tôi vẫn kiểm tra và xử lý các giá trị thiếu và dữ liệu ngoại lai nếu có. Các giá trị thiếu được thay thế bằng giá trị trung bình hoặc giá trị phổ biến nhất tùy theo loại biến.

**4. Chuẩn hóa và chuẩn bị dữ liệu:**

- Các biến số được chuẩn hóa để đảm bảo rằng tất cả các biến đều có thang đo tương đương, điều này giúp cải thiện hiệu suất của các thuật toán học máy.

- Dữ liệu được chia thành tập huấn luyện và tập kiểm tra để đánh giá hiệu quả của mô hình

Các bước trên đảm bảo rằng dữ liệu được chuẩn bị một cách kỹ lưỡng trước khi xây dựng mô hình dự đoán, giúp nâng cao độ chính xác và hiệu quả của các mô hình học máy được áp dụng.

**2.3. Chia dữ liệu thành tập huấn luyện và tập kiểm tra**

Dữ liệu được chia thành hai phần: tập huấn luyện (70%) và tập kiểm tra (30%) để đảm bảo mô hình được đánh giá khách quan. Việc chia dữ liệu theo tỉ lệ này giúp đảm bảo rằng mô hình không bị overfitting và có thể tổng quát hóa tốt hơn khi áp dụng trên dữ liệu mới.

Cụ thể, sau khi chia dữ liệu, tập huấn luyện có kích thước 318,780 mẫu và tập kiểm tra có kích thước 136,621 mẫu. Điều này đảm bảo tính cân bằng và đủ lớn cho cả quá trình huấn luyện và kiểm tra mô hình. Quá trình chia dữ liệu được thực hiện bằng cách sử dụng hàm train\_test\_split từ thư viện sklearn.

**CHƯƠNG III**

**PHƯƠNG PHÁP DỰ ĐOÁN**

**3.1. Các thuật toán sử dụng**

Trong nghiên cứu này, chúng tôi sử dụng thuật toán Gaussian Naive Bayes (GaussianNB) để dự đoán xác suất khách hàng tiềm năng. GaussianNB là một phương pháp phân loại xác suất dựa trên Định lý Bayes với giả định mạnh mẽ về tính độc lập của các đặc trưng (features). Đặc điểm nổi bật của GaussianNB bao gồm khả năng xử lý nhanh chóng các bài toán phân loại và hiệu suất cao trong nhiều tình huống khác nhau.

**3.1.1. Giới thiệu về Gaussian Naive Bayes**

Gaussian Naive Bayes là một thuật toán học máy thuộc họ Naive Bayes, được biết đến với các đặc điểm sau:

- *Đơn giản và hiệu quả*: GaussianNB dễ dàng triển khai và tính toán nhanh chóng ngay cả với các tập dữ liệu lớn, nhờ vào các giả định đơn giản hóa.

- *Hiệu suất cao trong nhiều trường hợp*: Mặc dù giả định về tính độc lập hoàn toàn giữa các đặc trưng có thể không chính xác trong thực tế, GaussianNB vẫn thường cho kết quả tốt và cạnh tranh trong nhiều bài toán phân loại.

- *Xử lý phân phối chuẩn*: GaussianNB giả định rằng các đặc trưng liên tục tuân theo phân phối chuẩn (Gaussian), do đó, nó đặc biệt phù hợp cho các tập dữ liệu có các đặc trưng phân phối xấp xỉ chuẩn.

**3.1.2. Quy trình huấn luyện và kiểm thử mô hình**

Quy trình xây dựng và đánh giá mô hình dự đoán bằng GaussianNB trong nghiên cứu này bao gồm các bước sau:

**1. Chia dữ liệu:**

Dữ liệu được chia thành tập huấn luyện và tập kiểm tra với tỷ lệ 70%\_30% tương ứng. Cụ thể, tập huấn luyện gồm 318,780 mẫu và tập kiểm tra gồm 136,621 mẫu. Việc chia dữ liệu này nhằm đảm bảo mô hình được đánh giá khách quan và tránh hiện tượng overfitting.

**2. Huấn luyện mô hình:**

- Sử dụng tập huấn luyện, mô hình GaussianNB được huấn luyện để học mối quan hệ giữa các đặc trưng và biến mục tiêu (khách hàng đặt hàng hay không đặt hàng). Quá trình huấn luyện bao gồm việc tính toán các thông số xác suất cần thiết cho Định lý Bayes, dựa trên các giả định về tính độc lập và phân phối chuẩn của các đặc trưng.

**3. Dự đoán và đánh giá:**

- Sau khi huấn luyện, mô hình được sử dụng để dự đoán trên tập test. Kết quả dự đoán được so sánh với giá trị thực tế để đánh giá hiệu quả của mô hình. Các chỉ số đánh giá bao gồm ma trận nhầm lẫn (confusion matrix) và độ chính xác (accuracy score).

**3.1.3. Phân tích kết quả:**

Kết quả từ tập kiểm tra cho thấy:

- Confusion Matrix thể hiện như sau:

[[129357, 1551],

[72, 5641]]

Điều này cho thấy mô hình đã phân loại chính xác 129,357 trường hợp không đặt hàng và 5,641 trường hợp đặt hàng. Tuy nhiên, cũng có 1,551 trường hợp không đặt hàng bị phân loại nhầm là đặt hàng và 72 trường hợp đặt hàng bị phân loại nhầm là không đặt hàng.

- Độ chính xác (accuracy score) đạt 0.9881204207259499, tương đương với 98.81%. Điều này chứng tỏ mô hình GaussianNB hoạt động hiệu quả trong việc dự đoán xác suất khách hàng tiềm năng.

**Kết luận**: kết quả này cho thấy mô hình có khả năng phân loại đúng phần lớn các khách hàng, giúp doanh nghiệp có thể tập trung vào các khách hàng tiềm năng cao và tối ưu hóa các chiến lược marketing và bán hàng. Trong tương lai, để cải thiện thêm hiệu suất của mô hình, có thể xem xét các phương pháp tiên tiến hơn như kết hợp nhiều thuật toán (ensemble methods), sử dụng các kỹ thuật tăng cường (boosting), hoặc thử nghiệm các thuật toán khác như Decision Trees, Random Forests, hoặc các mô hình dựa trên Deep Learning.

**3.2. Hàm predict\_proba**

Hàm predict\_proba trong mô hình Gaussian Naive Bayes được sử dụng để tính xác suất của từng khách hàng tiềm năng thuộc về nhóm khách hàng mục tiêu. Khi áp dụng hàm này, mô hình sẽ trả về xác suất mà một khách hàng cụ thể sẽ đặt hàng dựa trên các đặc trưng của họ. Quá trình này cho phép doanh nghiệp có được một cái nhìn sâu sắc và định lượng về khả năng chuyển đổi của từng khách hàng.

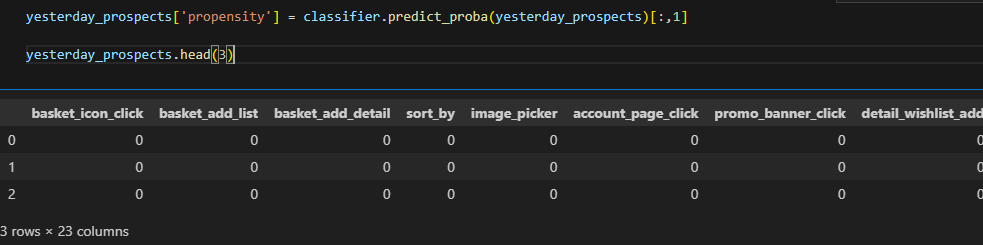
Trong nghiên cứu này, sau khi huấn luyện mô hình với tập dữ liệu huấn luyện, chúng tôi đã áp dụng mô hình này để dự đoán xác suất chuyển đổi của các khách hàng tiềm năng mới trong ngày hôm qua. Các bước thực hiện bao gồm:

1. **Tiền xử lý dữ liệu:**

Chúng tôi loại bỏ các cột không cần thiết như 'ordered', 'UserID', và 'device\_mobile' từ tập dữ liệu của các khách hàng tiềm năng. Điều này nhằm đảm bảo rằng chỉ các đặc trưng có liên quan mới được sử dụng trong quá trình dự đoán.

1. **Dự đoán xác suất chuyển đổi:**

Sử dụng hàm predict\_proba của mô hình GaussianNB, chúng tôi tính toán xác suất chuyển đổi của từng khách hàng tiềm năng. Hàm predict\_proba trả về một mảng xác suất cho mỗi khách hàng, trong đó xác suất của lớp dương (khách hàng đặt hàng) được sử dụng để đánh giá khả năng chuyển đổi.



1. **Kết quả dự đoán:**

Dữ liệu dự đoán được lưu trữ với xác suất chuyển đổi (propensity) của từng khách hàng. Điều này giúp doanh nghiệp xác định các khách hàng tiềm năng có xác suất cao nhất để tập trung các nỗ lực marketing và bán hàng.

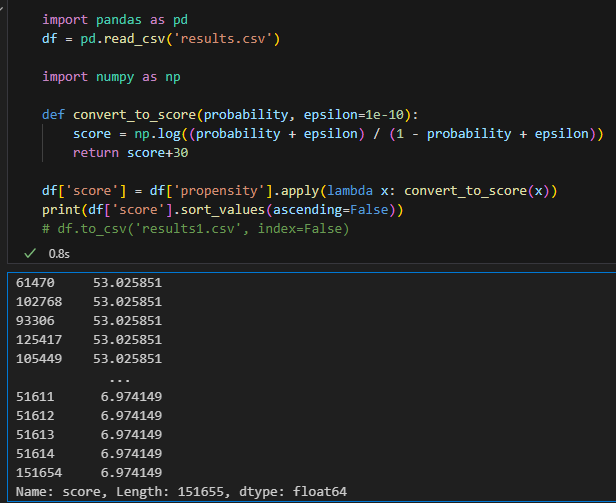
Kết quả cho thấy rằng các khách hàng trong tập dữ liệu kiểm tra cho ra kết quả khá ổn dựa trên các đặc trưng đã chọn thấp nhất là 0 và cao nhất là khoảng 53 tính đến thời điểm viết

Bằng cách sử dụng hàm predict\_proba, doanh nghiệp có thể phân tích sâu hơn về từng khách hàng, đưa ra các quyết định chiến lược dựa trên dữ liệu xác suất. Các khách hàng có xác suất chuyển đổi cao có thể được ưu tiên trong các chiến dịch tiếp thị và bán hàng, giúp tối ưu hóa tài nguyên và nâng cao hiệu quả kinh doanh. Đây là một bước tiến quan trọng trong việc áp dụng các kỹ thuật học máy để cải thiện quy trình quản lý khách hàng và tăng cường khả năng cạnh tranh của doanh nghiệp.

**3.3. Quy đổi xác suất thành điểm**

Để chuyển đổi xác suất thành điểm cụ thể, chúng tôi sử dụng công thức:

Trong đó, **p** là xác suất và **epsilon** là một giá trị rất nhỏ để tránh chia cho 0. Công thức này giúp chuẩn hóa các xác suất thành điểm số, dễ dàng hơn trong việc so sánh và phân loại khách hàng.



**CHƯƠNG IV**

**ỨNG DỤNG THỰC TẾ**

**4.1. Xác định các tương tác quan trọng**

Trong việc phân tích dữ liệu để hiểu rõ hành vi mua hàng của khách hàng, một bước quan trọng là xác định các tương tác quan trọng nhất. Những tương tác này có thể bao gồm nhiều hành vi khác nhau của người dùng trên trang web. Dưới đây là một số ví dụ tiêu biểu:

- *Xem một sản phẩm cụ thể*: Khi khách hàng xem chi tiết một sản phẩm, điều này có thể cho thấy sự quan tâm đặc biệt đến sản phẩm đó.

-*Thực hiện tìm kiếm*: Hành động tìm kiếm sản phẩm hoặc thông tin trên trang web thường phản ánh nhu cầu cụ thể và tức thời của khách hàng.

- *Thăm lại trang web sau khi đã rời đi*: Khi khách hàng quay lại trang web sau một thời gian, điều này có thể cho thấy họ vẫn còn quan tâm hoặc đang cân nhắc về một sản phẩm hoặc dịch vụ nào đó.

- *Đăng ký nhận bản tin hoặc các chương trình khuyến mãi*: Hành động này thể hiện sự quan tâm liên tục và mong muốn cập nhật thông tin mới nhất từ doanh nghiệp.

Việc xác định chính xác các tương tác quan trọng này giúp doanh nghiệp hiểu rõ hơn về hành vi của khách hàng và tạo ra các chiến lược tiếp thị chính xác và hiệu quả hơn.

**4.2. Chấm điểm khách hàng tiềm năng**

Sau khi đã xác định được các tương tác quan trọng, bước tiếp theo là sử dụng mô hình đã huấn luyện để đánh giá và chấm điểm các khách hàng dựa trên hành vi của họ trên trang web. Quá trình này bao gồm các bước sau:

- *Chấm điểm hàng ngày*: Các khách hàng tiềm năng cần được chấm điểm hàng ngày để đảm bảo rằng thông tin và hành vi mới nhất của họ được cập nhật liên tục. Điều này đặc biệt quan trọng đối với những khách hàng có hoạt động gần đây, vì họ có thể đang ở giai đoạn quan trọng trong hành trình mua hàng.

- *Đánh giá giá trị tiềm năng*: Các khách hàng được chấm điểm cao nhất sẽ được xác định là những khách hàng tiềm năng có giá trị nhất. Những khách hàng này sẽ là mục tiêu chính cho các chiến lược tái mục tiêu, vì họ có khả năng chuyển đổi thành khách hàng thực sự cao nhất.

**4.3. Tối ưu hóa chiến lược tái mục tiêu**

Dựa trên thông tin thu thập được từ quá trình chấm điểm khách hàng tiềm năng, doanh nghiệp có thể tiến hành tối ưu hóa chiến lược tái mục tiêu. Cụ thể, các bước sau đây cần được thực hiện:

- *Tinh chỉnh quảng cáo và thông điệp*: Sử dụng dữ liệu về sở thích và hành vi của khách hàng để điều chỉnh nội dung quảng cáo và thông điệp tiếp thị. Điều này giúp tăng khả năng thu hút và giữ chân khách hàng, cũng như nâng cao tỷ lệ chuyển đổi.

- *Điều chỉnh chiến lược marketing*: Đánh giá và tối ưu hóa các chiến lược marketing hiện tại để đảm bảo hiệu quả chi phí. Điều này có thể bao gồm việc phân bổ ngân sách quảng cáo một cách hợp lý, lựa chọn các kênh tiếp thị phù hợp nhất, và tối ưu hóa thời gian phát hành các chiến dịch.

- *Tăng cơ hội mua hàng*: Để thúc đẩy khách hàng tiềm năng thực hiện mua hàng, doanh nghiệp có thể cung cấp các ưu đãi hoặc khuyến mãi phù hợp. Các ưu đãi này cần được cá nhân hóa dựa trên hành vi và sở thích của khách hàng để đảm bảo tính hiệu quả cao nhất.

Nhờ việc áp dụng các phương pháp trên, doanh nghiệp không chỉ có thể tăng cường hiệu quả của chiến lược tiếp thị mà còn xây dựng được mối quan hệ lâu dài và bền vững với khách hàng.

**CHƯƠNG V**

**PINELINE**

**5.1 Các Bước và Tính Năng của Pipeline**

Pipeline là một hệ thống quan trọng trong việc tạo và cập nhật cơ sở dữ liệu người dùng hàng ngày. Để tối ưu hóa quá trình này, chúng ta cần thực hiện các bước và tích hợp các tính năng sau:

**1. Tạo và Cập Nhật Database Người Dùng:**

- *Tự động Thu thập và Cập nhật Thông tin Người Dùng theo ngày:* Việc cập nhật dữ liệu này nên được thực hiện hàng ngày để đảm bảo rằng thông tin luôn mới nhất và chính xác.

- *Tích hợp Dữ liệu Người Dùng vào Quy trình Pipeline*: Sau khi thu thập, dữ liệu người dùng cần được tích hợp vào quy trình pipeline. Điều này giúp cá nhân hóa các dự đoán và tối ưu hóa hiệu suất của mô hình. Ví dụ, dữ liệu về hành vi mua hàng, sở thích cá nhân, và các hoạt động trực tuyến của người dùng có thể được sử dụng để cải thiện độ chính xác của các mô hình dự đoán và chiến lược tiếp thị.

**5.2 Đánh Giá Pipeline**

Pipeline cần được đánh giá dựa trên một số tiêu chí quan trọng để đảm bảo hiệu quả và an toàn trong quá trình hoạt động:

**1. Hiệu Suất:**

- *Tự động Hóa*: Khả năng tự động hóa trong việc tạo và cập nhật cơ sở dữ liệu là yếu tố then chốt. Pipeline cần phải có khả năng xử lý khối lượng lớn dữ liệu mà không cần sự can thiệp thủ công nhiều.

- *Tích Hợp Hiệu Quả*: Dữ liệu người dùng cần được tích hợp một cách hiệu quả vào quy trình pipeline để đảm bảo rằng các mô hình dự đoán có thể sử dụng dữ liệu này một cách tối ưu.

**2. Độ Tin Cậy:**

- *Thu Thập và Cập Nhật Chính Xác***:** Khả năng thu thập và cập nhật thông tin chính xác từ nhiều nguồn khác nhau trên web là rất quan trọng. Pipeline phải đảm bảo rằng dữ liệu được cập nhật một cách kịp thời và không có sai sót.

**3. Tính Linh Hoạt:**

- *Điều Chỉnh và Mở Rộng*: Pipeline cần có khả năng điều chỉnh và mở rộng dễ dàng để tích hợp thêm các tính năng mới hoặc nguồn dữ liệu khác. Điều này bao gồm việc thêm các nguồn dữ liệu mới mà không làm gián đoạn quy trình hiện tại.

**4. Bảo Mật:**

- *Bảo Vệ Thông Tin Người Dùng*: Bảo mật là một yếu tố không thể thiếu. Pipeline phải tích hợp các cơ chế bảo mật mạnh mẽ để đảm bảo rằng thông tin người dùng được bảo vệ khỏi các mối đe dọa an ninh mạng và xâm nhập trái phép. Điều này bao gồm việc mã hóa dữ liệu và kiểm soát truy cập nghiêm ngặt.

Nhờ vào việc thực hiện các bước và tích hợp các tính năng trên, pipeline sẽ trở thành một công cụ mạnh mẽ giúp doanh nghiệp quản lý và tối ưu hóa dữ liệu người dùng một cách hiệu quả và an toàn.

## 

## 

## 

## 

## 

## 

## 

## 

## 

## 

**CHƯƠNG IV  
DEMO QUẢN LÝ KHÁCH HÀNG**

Chương này mô tả chi tiết các chức năng quản lý khách hàng được triển khai trong ứng dụng web Django, dựa trên nội dung mã nguồn bạn đã cung cấp.

**6.1. Hệ thống đăng ký người dùng:**

Hệ thống cung cấp chức năng register cho phép người dùng tạo tài khoản mới. Người dùng cần cung cấp thông tin cá nhân như tên, email, mật khẩu và xác nhận mật khẩu.

Hệ thống sẽ thực hiện các bước sau để xử lý đăng ký:

*-Kiểm tra thông tin đầu vào*: Hệ thống sẽ kiểm tra xem thông tin đăng ký có hợp lệ hay không, bao gồm:

-Kiểm tra độ dài và định dạng của tên người dùng.

-Đảm bảo email chưa được sử dụng bởi tài khoản khác.

-Xác nhận mật khẩu khớp với mật khẩu đã nhập.

-*Lưu trữ thông tin*: Nếu thông tin đăng ký hợp lệ, hệ thống sẽ lưu trữ thông tin người dùng vào cơ sở dữ liệu.

-*Kích hoạt tài khoản*: Hệ thống có thể gửi email kích hoạt đến người dùng để xác nhận tài khoản và hoàn tất quy trình đăng ký.

## 6.2. Bảng điều khiển quản lý(Dashboard):

**6.2.1 Tính Toán Điểm Tối Đa và Tối Thiểu:**

* Khởi tạo max\_score và min\_score dựa trên simulation đầu tiên nếu có simulation tồn tại.
* Lặp qua các simulation để xác định điểm tối đa và tối thiểu thực sự.

**6.2.2 Tính Tổng Số Người Dùng và số lượng khách hàng tiềm năng:**

**6.2.3 Phân tích điểm số khách hàng tiềm năng trong các khoảng :**

Hiển thị tỷ lệ phần trăm bản ghi mô phỏng thuộc vào các mức điểm khác nhau (Mức 1, Mức 2, Mức 3).

Có thể biểu diễn dữ liệu dưới dạng:

-Biểu đồ hình tròn (pie chart) để trực quan hóa tỷ lệ phần trăm của từng mức điểm.

-Biểu đồ thanh (bar chart) để so sánh tỷ lệ phần trăm giữa các mức điểm.

-Bảng dữ liệu hiển thị số lượng và tỷ lệ phần trăm cụ thể cho từng mức điểm.

**6.2.4 Biểu đồ trực quan:**

Hiển thị dữ liệu dưới dạng biểu đồ hoặc đồ thị để giúp người dùng dễ dàng nắm bắt xu hướng và phân tích thông tin.

Một số ví dụ về biểu đồ có thể sử dụng:

-Biểu đồ xu hướng (line chart) để hiển thị sự thay đổi của điểm số theo thời gian.

-Biểu đồ phân tán (scatter plot) để thể hiện mối quan hệ giữa các biến trong dữ liệu mô phỏng.

-Biểu đồ hộp (box plot) để so sánh phân bố điểm số giữa các nhóm dữ liệu khác nhau.

**6.2.5 Tương tác với dữ liệu:**

Cho phép người dùng tương tác với dữ liệu được hiển thị trên bảng điều khiển.

**Ví dụ:**

-Di chuột qua các điểm dữ liệu trên biểu đồ để hiển thị thông tin chi tiết.

-Nhấp vào các phân đoạn trong biểu đồ hình tròn để lọc dữ liệu theo mức điểm tương ứng.

-Sử dụng các bộ lọc để thu hẹp phạm vi dữ liệu hiển thị..

**6.3. Tải dữ liệu mô phỏng:**

Chức năng fileupload cho phép người dùng tải lên tệp CSV chứa dữ liệu mô phỏng. Hệ thống hỗ trợ tải lên tệp có dung lượng lớn và có thể xử lý nhiều bản ghi cùng lúc và có thể chia thành nhiều luồng xử lý.

Quy trình tải dữ liệu bao gồm các bước sau:

-Kiểm tra tệp: Hệ thống sẽ kiểm tra xem tệp được tải lên có hợp lệ hay không, bao gồm định dạng tệp (CSV), kích thước tệp và cấu trúc dữ liệu.

-Xử lý dữ liệu: Hệ thống sẽ đọc dữ liệu từ tệp CSV và chuyển đổi thành các định dạng phù hợp để lưu trữ trong cơ sở dữ liệu.

-Lưu trữ dữ liệu: Dữ liệu mô phỏng được lưu trữ vào cơ sở dữ liệu một cách an toàn và hiệu quả.

-Thông báo kết quả: Hệ thống sẽ thông báo cho người dùng về kết quả tải dữ liệu, bao gồm số lượng bản ghi được xử lý thành công và bất kỳ lỗi nào gặp phải trong quá trình xử lý.

**6.4. Mô phỏng tình huống(Simulation)**:

Chức năng simulation cho phép người dùng mô phỏng các tình huống dựa trên dữ liệu đã được tải lên.

Người dùng cần cung cấp các thông tin sau để thực hiện mô phỏng:

-Chọn thuộc tính: Lựa chọn các thuộc tính (tính năng) có liên quan đến tình huống mô phỏng.

-Cung cấp giá trị đầu vào: Cung cấp giá trị cho các thuộc tính đã chọn.

Hệ thống sẽ thực hiện các bước sau để thực hiện mô phỏng:

-Tải mô hình đã được đào tạo: Hệ thống sẽ tải mô hình học máy đã được đào tạo trước đó để dự đoán kết quả cho tình huống mô phỏng.

-Chuẩn bị dữ liệu đầu vào: Dữ liệu đầu vào được chuẩn bị theo định dạng phù hợp với mô hình học máy.

-Thực hiện dự đoán: Mô hình học máy sẽ dự đoán kết quả cho tình huống mô phỏng dựa trên dữ liệu đầu vào.

-Tính toán điểm số: Hệ thống sẽ tính toán điểm số dựa trên kết quả dự đoán và các tiêu chí đánh giá được định nghĩa.

-Lưu trữ kết quả mô phỏng: Kết quả mô phỏng, bao gồm dữ liệu đầu vào, điểm số và các thông tin khác, được lưu trữ vào cơ sở dữ liệu.

## 6.5. DataBase

**6.5.1 Lưu trữ dữ liệu mô phỏng**:

Hệ thống cung cấp chức năng save\_file cho phép người dùng lưu trữ dữ liệu mô phỏng đã được lọc hoặc xử lý theo các tiêu chí cụ thể sang tệp CSV.

Quy trình lưu trữ dữ liệu bao gồm các bước sau:

-Lấy dữ liệu: Hệ thống sẽ truy xuất dữ liệu mô phỏng từ cơ sở dữ liệu dựa trên các tiêu chí lọc được người dùng lựa chọn.

-Chuẩn bị dữ liệu: Dữ liệu được chuẩn bị theo định dạng phù hợp với tệp CSV, bao gồm tiêu đề cột và dữ liệu hàng.

-Tạo tệp CSV: Hệ thống tạo tệp CSV mới và ghi dữ liệu đã chuẩn bị vào tệp.

-Cung cấp tệp cho người dùng: Tệp CSV được tạo sẽ được cung cấp cho người dùng để tải xuống.

**6.5.2 Quản lý cơ sở dữ liệu:**

Chức năng database cung cấp giao diện để người dùng quản lý dữ liệu mô phỏng được lưu trữ trong cơ sở dữ liệu.

Người dùng có thể thực hiện các thao tác sau:

-Xem danh sách bản ghi: Hiển thị danh sách tất cả các bản ghi mô phỏng, bao gồm thông tin về điểm số, thời gian tạo, v.v.

-Lọc dữ liệu: Lọc dữ liệu theo các tiêu chí như điểm số, khoảng thời gian, hoặc các thuộc tính khác.

-Phân trang kết quả: Phân trang danh sách bản ghi để dễ dàng quản lý với số lượng lớn dữ liệu.

-Xem chi tiết bản ghi: Xem chi tiết thông tin của một bản ghi mô phỏng cụ thể.

-Xóa bản ghi: Xóa một hoặc nhiều bản ghi mô phỏng khỏi cơ sở dữ liệu.

**6.5.3 Quản lý người dùng:**

Hệ thống cung cấp các chức năng quản lý người dùng, bao gồm:

- users: Hiển thị danh sách tất cả người dùng đã đăng ký, bao gồm thông tin về user\_id và các hoạt động trên trang web.

- Mô tả chi tiết một khách hàng về số điểm, hạng của khách hàng đó, rank đạt được

-user\_delete: Xóa một người dùng cụ thể khỏi hệ thống.

- Có thể chỉnh sửa các thao tác trên web của 1 người nào đó

-changePassword: Cho phép người dùng thay đổi mật khẩu tài khoản của họ.

-Lọc ra những khách hàng tiềm năng level1, level2, level3

-Filter: để lọc ra những cột mà mình muốn xem

**6.6. Xử lý lỗi:**

Hệ thống sử dụng các phương pháp xử lý lỗi để đảm bảo hoạt động ổn định và cung cấp trải nghiệm tốt cho người dùng.

Các lỗi phổ biến được xử lý bao gồm:

-Lỗi nhập dữ liệu: Hệ thống hiển thị thông báo lỗi khi người dùng nhập dữ liệu không hợp lệ.

-Lỗi tải tệp: Hệ thống báo lỗi khi tệp tải lên không hợp lệ hoặc gặp sự cố trong quá trình xử lý.

-Lỗi cơ sở dữ liệu: Hệ thống ghi lại lỗi truy cập cơ sở dữ liệu và thực hiện các biện pháp xử lý phù hợp.

-Lỗi hệ thống: Hệ thống ghi lại lỗi hệ thống và thông báo cho quản trị viên để khắc phục.

**CHƯƠNG VII**

**KẾT LUẬN**

**7.1 Tóm Tắt Kết Quả**

Trong báo cáo này, chúng ta đã trình bày chi tiết về phương pháp và kết quả của việc dự đoán điểm của khách hàng tiềm năng. Qua các bước phân tích và ứng dụng mô hình, kết quả cho thấy mô hình dự đoán của chúng ta có khả năng dự đoán chính xác cao. Điều này giúp doanh nghiệp quản lý khách hàng hiệu quả hơn, tập trung vào những khách hàng tiềm năng nhất và tối ưu hóa các chiến lược tiếp thị và bán hàng.

**7.2 Hướng Phát Triển Trong Tương Lai**

Để tiếp tục cải thiện và mở rộng khả năng của mô hình, chúng ta có thể hướng tới một số phát triển sau:

**1. Sử Dụng Các Kỹ Thuật Học Sâu:**

- *Mở Rộng Ứng Dụng Mô Hình Học Sâu*: Việc áp dụng các mô hình học sâu như mạng nơ-ron sâu (Deep Neural Networks), các mô hình tự học (Self-Learning Models), và mạng nơ-ron tái cấu trúc (Recurrent Neural Networks) có thể giúp tối ưu hóa quá trình dự đoán. Những mô hình này có khả năng học hỏi và thích ứng với các đặc điểm phức tạp hơn trong dữ liệu, từ đó cung cấp các dự đoán chính xác và cá nhân hóa hơn cho từng khách hàng.

**2. Tăng Cường Dữ Liệu:**

- *Mở Rộng Quy Mô Dữ Liệu*: Việc tích hợp thêm thông tin từ nhiều nguồn khác nhau, bao gồm dữ liệu từ mạng xã hội, dữ liệu địa lý, dữ liệu giao dịch, và dữ liệu sản phẩm, sẽ giúp cải thiện độ chính xác của các dự đoán. Dữ liệu đa dạng và phong phú hơn sẽ cung cấp cái nhìn toàn diện về hành vi và sở thích của khách hàng, từ đó giúp cá nhân hóa tốt hơn các chiến lược tiếp thị.

**3. Tích Hợp Các Nguồn Dữ Liệu Mới**:

- *Khám Phá và Tích Hợp Dữ Liệu Mới*: Để có cái nhìn chi tiết hơn về hành vi của khách hàng, chúng ta có thể tích hợp các nguồn dữ liệu mới như dữ liệu về xu hướng thị trường, dữ liệu từ thiết bị IoT (Internet of Things), và dữ liệu từ các ứng dụng di động. Những nguồn dữ liệu này sẽ cung cấp thêm thông tin giá trị, giúp mô hình dự đoán trở nên toàn diện hơn và phù hợp hơn với thực tế.

**4. Cải Tiến Liên Tục**:

- *Quy Trình Cập Nhật và Kiểm Định Định Kỳ*: Thiết lập một quy trình để cập nhật và kiểm định mô hình định kỳ là rất quan trọng. Việc theo dõi hiệu suất và hiệu quả của hệ thống, cùng với việc áp dụng các phản hồi từ người dùng và dữ liệu mới, sẽ giúp cải thiện mô hình liên tục. Điều này đảm bảo rằng mô hình luôn phù hợp với những thay đổi trong hành vi và nhu cầu của khách hàng, đồng thời điều chỉnh chiến lược marketing và bán hàng một cách kịp thời và chính xác.

Việc thực hiện các hướng phát triển này không chỉ giúp cải thiện độ chính xác và hiệu quả của mô hình dự đoán mà còn tăng cường khả năng cá nhân hóa và tối ưu hóa chiến lược tiếp thị, từ đó nâng cao trải nghiệm khách hàng và gia tăng doanh thu cho doanh nghiệp.

**LÀM VIỆC NHÓM**

Nhóm làm việc theo 2 hình thức online và offline

Phân chia công việc:

* Nhơn: Điều hành nhóm đưa ra ý tưởng làm việc, phân chia công việc, chạy airflow, deploy docker, phụ trách report và slide
* Diện: Code web demo trang admin quản lý khách hàng tiềm năng bằng Django
* Đạt: Phụ trách làm slide, đưa ra các đặc trưng quan trọng trong bộ dữ liệu, hỗ trợ những việc khác
* Huy: Phụ trách làm báo cáo, tổng quan lại kiến thức đã làm, hỗ trợ những công việc khác.
* Thọ: Phụ trách làm báo cáo, đưa ra ý tưởng để cải thiện web, hỗ trợ những công việc khác.

Tổng số lần gặp nhau: 10

Tổng thời gian gặp nhau: 52 giờ