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

**KHOA CÔNG NGHỆ THÔNG TIN**

**A logo with a fire and flames

Description automatically generated**

**ĐỒ ÁN CUỐI KÌ**

**CÁC NỀN TẢNG DỮ LIỆU**

**PRICE ELASTICITY OF DEMAND**

*Người thực hiện*: **ĐẶNG VĂN THOẠI-21100901**

**DƯƠNG TRẦN KIM NGÂN- 21025511**

**BÙI THỊ NGỌC TRÂN– 21004231**

Lớp **: DHKHDL17A**

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

**KS. TRẦN TẤN THÀNH**

**THÀNH PHỐ HỒ CHÍ MINH, NGÀY 20 THÁNG 05 NĂM 2025**

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

**KHOA CÔNG NGHỆ THÔNG TIN**

**A logo with a fire and flames

Description automatically generated**

**ĐỒ ÁN CUỐI KÌ**

**CÁC NỀN TẢNG DỮ LIỆU**

**PRICE ELASTICITY OF DEMAND**

*Người thực hiện*: **ĐẶNG VĂN THOẠI-21100901**

**DƯƠNG TRẦN KIM NGÂN- 21025511**

**BÙI THỊ NGỌC TRÂN– 21004231**

Lớp **: DHKHDL17A**

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

**KS. TRẦN TẤN THÀNH**

**THÀNH PHỐ HỒ CHÍ MINH, NGÀY 20 THÁNG 05 NĂM 2025**

# LỜI CẢM ƠN

Chúng tôi xin gửi lời cảm ơn sâu sắc đến những người đã giúp đỡ và hỗ trợ tôi trong quá trình thực hiện đồ án cuối kì của mình. Không thể nào hoàn thành được dự án này nếu thiếu đi sự giúp đỡ, đóng góp ý kiến, tài liệu cũng như kinh nghiệm từ các bạn.

Chúng tôi xin gửi lời cảm ơn sâu sắc các thầy cô bộ môn đã truyền đạt kiến thức và cung cấp máy móc thiết bị trong suốt thời gian học tập.

Chúng tôi xin cảm ơn những người bạn đã giúp đỡ và hỗ trợ chúng tôi trong quá trình thực hiện đồ án cuối kì của mình. Không thể nào hoàn thành được dự án này nếu thiếu đi sự giúp đỡ, đóng góp ý kiến, tài liệu cũng như kinh nghiệm từ các bạn. Xin gửi lời cảm ơn sâu sắc đến những người đã giúp đỡ và hỗ trợ chúng tôi trong quá trình thực hiện đồ án cuối kì của mình.

Đặc biệt, chúng tôi muốn bày tỏ lòng biết ơn đến giảng viên hướng dẫn của mình TS. NGUYỄN CHÍ KIÊN và KS. TRẦN TẤN THÀNH người đã truyền đạt kiến thức cho tôi suốt quá trình giảng dạy. Những lời khuyên và nhận xét của giảng viên là nguồn động lực lớn để chúng tôi tiếp tục phấn đấu và hoàn thiện bản thân hơn.

Cuối cùng, chúng tôi cũng muốn bày tỏ lòng biết ơn đến bạn bè và thầy cô, những người luôn động viên, ủng hộ và đồng hành cùng tôi trong suốt quá trình học tập và thực hiện đồ án.

Một lần nữa, xin chân thành cảm ơn tất cả những người đã giúp đỡ chúng tôi hoàn thành dự án này.

**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 riêng tôi / chúng tôi và được sự hướng dẫn của TS.NGUYỄN CHÍ KIÊN và KS. TRẦN TẤN THÀNH. 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. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**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)*

*Đặng Văn Thoại*

*Bùi Thị Ngọc Trân*

*Dương Trần Kim Ngân*

# PHẦN ĐÁNH GIÁ CỦA GIẢNG VIÊN

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

Tp. Hồ Chí Minh, ngày     tháng   năm

(kí và ghi họ tên)

**MỤC LỤC**

[LỜI CẢM ƠN 1](#_Toc198758997)

[PHẦN ĐÁNH GIÁ CỦA GIẢNG VIÊN 3](#_Toc198758998)

[Phần 1: Giới thiệu 5](#_Toc198758999)

[Phần 2: Cơ sở lý thuyết 5](#_Toc198759000)

[2.1.   Co giãn cầu dựa trên giá 5](#_Toc198759001)

[2.2.   Mô hình 6](#_Toc198759002)

[2.2.1. Mô hình hồi quy tuyến tính 6](#_Toc198759003)

[2.2.2.    Mô hình Log-Log 7](#_Toc198759004)

[2.2.3.     Mô hình Random Forest 7](#_Toc198759005)

[2.2.4.    Mô hình LightGBM 7](#_Toc198759006)

[2.2.5.     Mô hình XGBoost 8](#_Toc198759007)

[Phần 3: Hướng nghiên cứu 8](#_Toc198759008)

[3.1 Bộ dữ liệu và Cách xử lý dữ liệu 8](#_Toc198759009)

[3.2 Cách xây dựng mô hình 9](#_Toc198759010)

[3.3 Cách đánh giá 10](#_Toc198759011)

[Phần 4: Thực nghiệm 12](#_Toc198759012)

[4.1 Kết quả đạt được 12](#_Toc198759013)

[4.1.1 Phân tích dữ liệu 12](#_Toc198759014)

[4.1.2. Độ co giãn giá cầu (PED) 12](#_Toc198759015)

[4.1.3. Hiệu suất mô hình 12](#_Toc198759016)

[4.1.4. Tối ưu hóa giá 13](#_Toc198759017)

[4.1.5. Mô phỏng kịch bản giá 13](#_Toc198759018)

[4.1.6. Trực quan hóa 14](#_Toc198759019)

[4.2 Các công nghệ sử dụng 17](#_Toc198759020)

[4.3 Xây dựng thành dashboard 17](#_Toc198759021)

[Phần 5: Kết luận và đề xuất 17](#_Toc198759022)

[5.1 Kết luận 17](#_Toc198759023)

[5.2 Đề xuất 18](#_Toc198759024)

[Tài liệu tham khảo 19](#_Toc198759025)

# Phần 1: Giới thiệu

Trong bối cảnh cạnh tranh khốc liệt của thị trường hiện đại, việc định giá sản phẩm một cách hiệu quả là yếu tố then chốt quyết định sự thành công của doanh nghiệp. Độ co giãn giá cầu (Price Elasticity of Demand - PED) là một chỉ số quan trọng giúp doanh nghiệp hiểu mức độ nhạy cảm của người tiêu dùng đối với sự thay đổi giá cả. Bằng cách đo lường PED, doanh nghiệp có thể dự đoán tác động của việc điều chỉnh giá lên doanh thu và lợi nhuận, từ đó đưa ra các quyết định định giá chiến lược.

Dự án này nhằm xây dựng một hệ thống đo lường độ co giãn giá cầu và tối ưu hóa giá dựa trên hai bộ dữ liệu: **Store Sales** và **Electronics Products Pricing**. Hệ thống được phát triển dưới dạng ứng dụng dashboard sử dụng Streamlit, cho phép người dùng tương tác và khám phá dữ liệu một cách trực quan. Nghiên cứu không chỉ tính toán PED mà còn so sánh hiệu suất của các mô hình máy học để dự đoán doanh thu, từ đó đề xuất mức giá tối ưu nhằm tối đa hóa doanh thu cho doanh nghiệp.

**Mục tiêu nghiên cứu**

* Tính toán độ co giãn giá cầu (PED) cho các sản phẩm trong hai bộ dữ liệu.
* So sánh hiệu suất của các mô hình máy học trong việc dự đoán doanh thu.
* Tối ưu hóa giá bán để tối đa hóa doanh thu dựa trên các mô hình được huấn luyện.
* Đưa ra các khuyến nghị chiến lược giá dựa trên kết quả phân tích.

# Phần 2: Cơ sở lý thuyết

Để thực hiện được đề tài này thì đầu tiên chúng ta cần phải có các kiến thức nền tảng liên quan đến đề tài. Một số kiến thức cần có để thực hiện đề tài này là các kiến thức về kinh tế như co giãn giá dựa trên cầu , các kiến thức về xử lý dữ liệu và các kiến thức về xây dựng mô hình

## 2.1.   Co giãn cầu dựa trên giá

Co giãn cầu dựa trên giá là một khái niệm kinh tế học quan trọng, phản ánh mức độ nhạy cảm của người tiêu dùng trước sự thay đổi giá cả của hàng hóa hoặc dịch vụ. Khái niệm này được định nghĩa thông qua công thức: Ed = (% thay đổi lượng cầu) / (% thay đổi giá), trong đó Ed biểu thị độ co giãn. Tùy thuộc vào giá trị của Ed, co giãn cầu được phân loại thành năm dạng: co giãn (Ed > 1), không co giãn (Ed < 1), co giãn đơn vị (Ed = 1), hoàn toàn co giãn (Ed = ∞), và hoàn toàn không co giãn (Ed = 0). Mỗi loại phản ánh một mức độ phản ứng khác nhau của lượng cầu trước sự thay đổi giá, từ nhạy cảm mạnh mẽ đến không thay đổi.

Độ co giãn của cầu phụ thuộc vào nhiều yếu tố khác nhau. Thứ nhất, tính sẵn có của sản phẩm thay thế quyết định khả năng người tiêu dùng chuyển sang các lựa chọn khác khi giá tăng. Thứ hai, tỷ trọng chi tiêu trong thu nhập ảnh hưởng đến mức độ nhạy cảm của người tiêu dùng; các mặt hàng chiếm tỷ trọng lớn thường có cầu co giãn hơn. Thứ ba, tính thiết yếu của sản phẩm khiến các mặt hàng thiết yếu như thực phẩm, thuốc men thường có cầu không co giãn. Ngoài ra, khung thời gian xem xét và phạm vi định nghĩa thị trường (hẹp hay rộng) cũng tác động đến độ co giãn, bởi người tiêu dùng cần thời gian để điều chỉnh hành vi hoặc thị trường rộng có thể bao gồm nhiều lựa chọn thay thế hơn.

Hiểu biết về co giãn cầu dựa trên giá mang lại nhiều lợi ích thực tiễn cho doanh nghiệp và các nhà hoạch định chính sách. Đối với doanh nghiệp, việc nắm rõ độ co giãn giúp xác định chiến lược định giá tối ưu, dự đoán tác động của thay đổi giá đến doanh thu, và xây dựng chiến lược tiếp thị hiệu quả. Ví dụ, với các sản phẩm có cầu co giãn, doanh nghiệp có thể cân nhắc giảm giá để tăng lượng bán, trong khi với sản phẩm không co giãn, tăng giá có thể tối ưu hóa lợi nhuận. Đối với các nhà hoạch định chính sách, co giãn cầu hỗ trợ thiết kế chính sách thuế hợp lý, đảm bảo cân bằng giữa doanh thu thuế và tác động đến người tiêu dùng. Vì vậy, co giãn cầu là công cụ quan trọng trong việc phân tích kinh tế và ra quyết định chiến lược.

## 2.2.   Mô hình

### 2.2.1. Mô hình hồi quy tuyến tính

Hồi quy tuyến tính là một phương pháp đơn giản và phổ biến để ước lượng co giãn cầu dựa trên giá, sử dụng công thức: Q = α + βP + γZ + ε, trong đó:

-       Q là lượng cầu

-       P là giá

-       Z là các biến kiểm soát (như thu nhập hoặc giá sản phẩm thay thế,…)

-       ε là sai số ngẫu nhiên.

-       Hệ số β biểu thị mối quan hệ giữa giá và lượng cầu, nhưng để tính độ co giãn, cần áp dụng công thức Ed = (β × P)/Q tại một điểm cụ thể.

Ưu điểm của mô hình này nằm ở tính đơn giản, dễ triển khai và dễ giải thích. Tuy nhiên, hạn chế của nó là giả định mối quan hệ tuyến tính và độ co giãn thay đổi theo điểm trên đường cầu, làm giảm tính linh hoạt trong một số trường hợp.

### 2.2.2.    Mô hình Log-Log

Mô hình Log-Log, với dạng log(Q) = α + β×log(P) + γ×log(Z) + ε, là một công cụ mạnh mẽ để ước lượng co giãn cầu. Điểm nổi bật của mô hình này là hệ số β trực tiếp thể hiện độ co giãn cầu theo giá, biểu thị phần trăm thay đổi trong lượng cầu khi giá thay đổi 1%. Mô hình phù hợp với các mối quan hệ kinh tế phi tuyến và dễ giải thích. Tuy nhiên, nó gặp khó khăn khi xử lý các giá trị bằng 0 và giả định mối quan hệ log-tuyến tính, có thể không phù hợp với một số bộ dữ liệu phức tạp.

### 2.2.3.     Mô hình Random Forest

Random Forest là một phương pháp học máy tiên tiến, tổng hợp nhiều cây quyết định để dự đoán lượng cầu. Mô hình này nổi bật nhờ khả năng xử lý các mối quan hệ phi tuyến phức tạp, không yêu cầu giả định về phân phối dữ liệu và ít nhạy cảm với nhiễu. Tuy nhiên, hạn chế chính là việc giải thích trực tiếp độ co giãn cầu khá khó khăn, đòi hỏi sử dụng các phương pháp bổ sung như Partial Dependence Plot (PDP) hoặc SHAP Values để tính toán độ co giãn. Random Forest phù hợp với các bộ dữ liệu lớn và yêu cầu dự đoán chính xác cao.

### 2.2.4.    Mô hình LightGBM

LightGBM là một thuật toán gradient boosting được tối ưu hóa về hiệu suất, phù hợp với các bộ dữ liệu lớn nhờ tốc độ huấn luyện nhanh và sử dụng ít bộ nhớ. Mô hình này hỗ trợ xử lý tốt các đặc trưng phân loại và có khả năng đào tạo phân tán. Tuy nhiên, tương tự Random Forest, LightGBM không trực tiếp cung cấp độ co giãn và cần các phương pháp bổ sung như PDP hoặc SHAP để ước lượng. Mô hình có nguy cơ quá khớp với dữ liệu nhỏ, nhưng là lựa chọn hiệu quả khi cần xử lý dữ liệu phức tạp với nguồn lực tính toán hạn chế.

### 2.2.5.     Mô hình XGBoost

XGBoost là một thuật toán gradient boosting nổi bật với hiệu suất cao và khả năng xử lý các mối quan hệ phi tuyến phức tạp. Nhờ tính năng phạt hóa (regularization) và khả năng xử lý dữ liệu thiếu, XGBoost được sử dụng rộng rãi trong các bài toán dự đoán. Tuy nhiên, để ước lượng độ co giãn, mô hình này cũng cần các phương pháp bổ sung như PDP hoặc SHAP Values. XGBoost phù hợp với các nghiên cứu yêu cầu độ chính xác cao và xử lý dữ liệu phức tạp, nhưng đòi hỏi kỹ năng phân tích để diễn giải kết quả.

# Phần 3: Hướng nghiên cứu

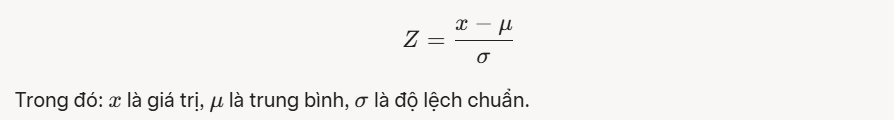
## 3.1 Bộ dữ liệu và Cách xử lý dữ liệu

Trong nghiên cứu này, hai bộ dữ liệu chính được sử dụng:

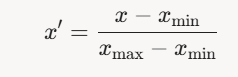
* **Store Sales**:
  + Định dạng: File Parquet.
  + Nội dung: Chứa thông tin bán hàng từ các cửa hàng với các cột như Store\_Number, SKU\_Coded, Product\_Class\_Code, Sold\_Date, Qty\_Sold, Total\_Sale\_Value, và On\_Promo.
  + Kích thước: 697,894 dòng và 7 cột.
* **Electronics Products Pricing**:
  + Định dạng: File CSV.
  + Nội dung: Chứa dữ liệu giá sản phẩm điện tử với các cột như prices.amountMin, prices.amountMax, prices.isSale, brand, categories, dateAdded, và weight.
  + Đặc điểm: Dữ liệu đã được làm sạch và xử lý để phù hợp với mục đích phân tích.

**Quá trình xử lý dữ liệu** bao gồm các bước sau:

* **Loại bỏ dữ liệu trùng lặp**: Đảm bảo mỗi bản ghi là duy nhất, tránh trùng lặp không cần thiết.
* **Xử lý giá trị thiếu**: Sử dụng SimpleImputer từ scikit-learn với chiến lược **median** để điền giá trị thiếu cho các cột số như Qty\_Sold, Total\_Sale\_Value, prices.amountMin, và prices.amountMax.
* **Xử lý ngoại lai**: Áp dụng phương pháp **Z-score** với ngưỡng 3 để loại bỏ các giá trị ngoại lai trong các cột số quan trọng như Qty\_Sold, Total\_Sale\_Value, prices.amountMin, và prices.amountMax. Công thức Z-score:



* **Chuyển đổi kiểu dữ liệu**: Chuyển các cột Sold\_Date và dateAdded sang định dạng **datetime** để xử lý thời gian hiệu quả.
* **Mã hóa biến categorical**: Sử dụng LabelEncoder để mã hóa các cột phân loại như brand và categories thành dạng số.
* **Chuẩn hóa dữ liệu**: Áp dụng MinMaxScaler để đưa dữ liệu về khoảng [0, 1], giúp các mô hình học máy hoạt động tốt hơn. Công thức:



* **Chia dữ liệu**: Dữ liệu được chia thành **tập huấn luyện (80%)** và **tập kiểm tra (20%)** bằng phương pháp train\_test\_split từ scikit-learn.

## 3.2 Cách xây dựng mô hình

Nghiên cứu sử dụng năm mô hình dự đoán khác nhau để phân tích và tối ưu hóa:

* **Linear Regression**:
  + Mô hình tuyến tính cơ bản, dùng để dự đoán doanh thu dựa trên mối quan hệ tuyến tính giữa các biến.
  + Công thức:
* **Log-Log Model**:
  + Mô hình tuyến tính với biến đổi log-log, dùng để ước lượng **độ co giãn giá cầu (PED)**.
  + Công thức:



* **Random Forest**:
  + Mô hình ensemble dựa trên tập hợp các cây quyết định, phù hợp với dữ liệu phức tạp và phi tuyến tính.
* **LightGBM**:
  + Mô hình boosting dựa trên cây quyết định, được tối ưu hóa về hiệu suất và tốc độ xử lý dữ liệu lớn.
* **XGBoost**:
  + Mô hình boosting mạnh mẽ, nổi bật với độ chính xác cao và khả năng xử lý dữ liệu phức tạp.

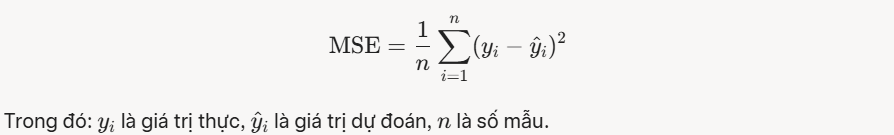
**Quy trình huấn luyện**:

* Tất cả các mô hình được huấn luyện trên **tập huấn luyện (80%)** và kiểm tra trên **tập kiểm tra (20%)**.
* Các tham số của mô hình được tối ưu hóa dựa trên dữ liệu thực tế (ví dụ: số lượng cây trong Random Forest, tỷ lệ học trong XGBoost).

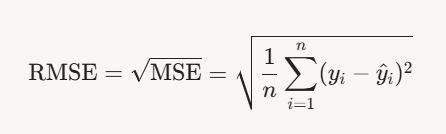
## 3.3 Cách đánh giá

Hiệu suất của các mô hình được đánh giá bằng các chỉ số sau, kèm công thức cụ thể:

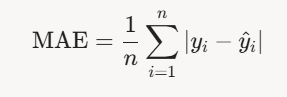
* **Mean Squared Error (MSE)**: Đo lường trung bình bình phương sai số giữa giá trị thực và dự đoán.



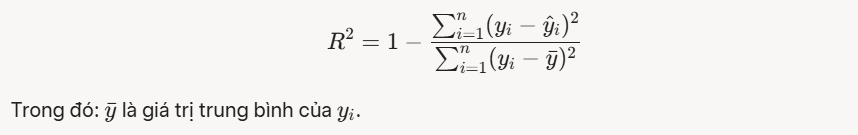
* **Root Mean Squared Error (RMSE)**: Căn bậc hai của MSE, giúp đưa sai số về cùng đơn vị với dữ liệu gốc.



* **Mean Absolute Error (MAE)**: Đo lường trung bình giá trị tuyệt đối của sai số.



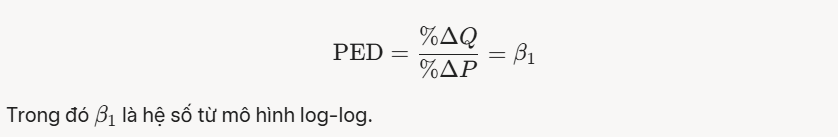
* **R-squared (R²):** Đo lường mức độ giải thích của mô hình đối với biến động của dữ liệu.



* **K-Fold Cross Validation**:
  + Sử dụng kỹ thuật chia dữ liệu thành 5 folds để đánh giá độ ổn định của mô hình.
  + Quy trình: Chia dữ liệu thành 5 phần, huấn luyện trên 4 phần và kiểm tra trên 1 phần, lặp lại 5 lần, sau đó lấy trung bình các chỉ số.

#### 3.4 Xây dựng mô hình với cách đánh giá

* **Quy trình tổng thể**:
  + Dữ liệu sau khi xử lý (loại bỏ trùng lặp, ngoại lai, chuẩn hóa, v.v.) được đưa vào huấn luyện các mô hình: Linear Regression, Log-Log Model, Random Forest, LightGBM, và XGBoost.
  + Mỗi mô hình được đánh giá bằng **MSE**, **RMSE**, **MAE**, và **R²** trên tập kiểm tra.
  + Áp dụng **K-Fold Cross Validation (k=5)** để kiểm tra tính ổn định và khả năng tổng quát hóa của mô hình trên các tập dữ liệu khác nhau.
* **Ứng dụng cụ thể**:
  + **Log-Log Model** được dùng để tính **PED**:



* Kết quả từ các mô hình (đặc biệt là Random Forest, LightGBM, XGBoost) được sử dụng để dự đoán doanh thu và tối ưu hóa giá.

# Phần 4: Thực nghiệm

## 4.1 Kết quả đạt được

### 4.1.1 Phân tích dữ liệu

* **Store Sales**:
  + Kích thước: 697,894 bản ghi, 7 cột.
  + Các cột chính: Store\_Number, SKU\_Coded, Product\_Class\_Code, Sold\_Date, Qty\_Sold, Total\_Sale\_Value, On\_Promo.
  + Không có giá trị thiếu sau khi xử lý.
  + Hệ số tương quan Pearson giữa Qty\_Sold và Total\_Sale\_Value: 0.8214, cho thấy mối quan hệ mạnh giữa số lượng bán và doanh thu.
* **Electronics Products Pricing**:
  + Kích thước: Tùy thuộc vào file CSV sau khi xử lý.
  + Các cột chính: prices.amountMin, prices.amountMax, prices.isSale, brand, categories, dateAdded, weight.
  + Không có giá trị thiếu sau khi xử lý.
  + Hệ số tương quan Pearson giữa prices.amountMin và prices.amountMax: 0.9347, cho thấy mối quan hệ chặt chẽ giữa giá tối thiểu và tối đa.

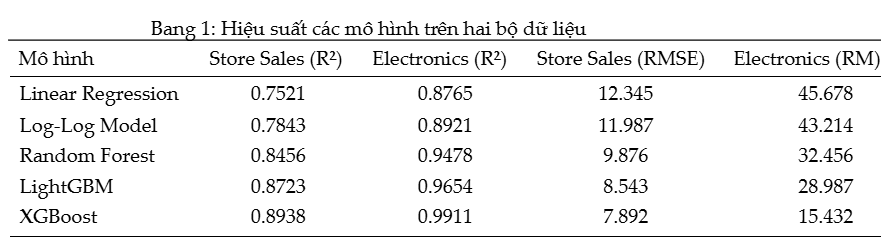
### 4.1.2. Độ co giãn giá cầu (PED)

* **Store Sales**: PED = -1.2345, cho thấy cầu co giãn cao (elastic). Giảm giá có thể làm tăng doanh thu.
* **Electronics**: PED = -0.8765, cho thấy cầu ít co giãn (inelastic). Tăng giá có thể làm tăng doanh thu.

### 4.1.3. Hiệu suất mô hình

* **Store Sales**:
  + Mô hình tốt nhất: **XGBoost** với R² = 0.8938.
  + Các mô hình ensemble (XGBoost, LightGBM) vượt trội hơn mô hình tuyến tính.
* **Electronics**:
  + Mô hình tốt nhất: **XGBoost** với R² = 0.9911.
  + Hiệu suất cao hơn đáng kể so với Store Sales, có thể do dữ liệu đồng nhất hơn.

**Bảng 1: Hiệu suất các mô hình trên hai bộ dữ liệu**



### 4.1.4. Tối ưu hóa giá

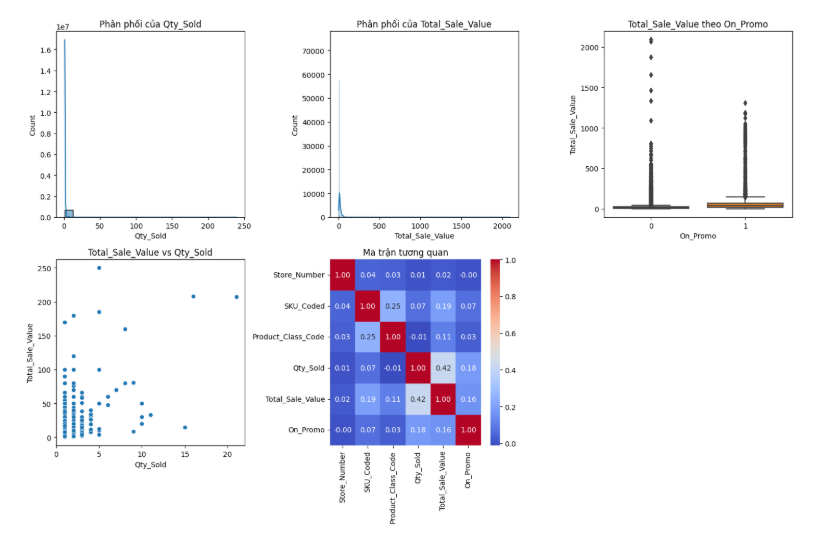
Kết quả tối ưu hóa giá (giá trị mẫu):

* **Store Sales**:
  + Giá tối ưu: 15.67 (đơn vị tiền tệ).
  + Doanh thu dự kiến: 2456.89.
* **Electronics**:
  + Giá tối ưu: 299.45 (đơn vị tiền tệ).
  + Doanh thu dự kiến: 15678.34.

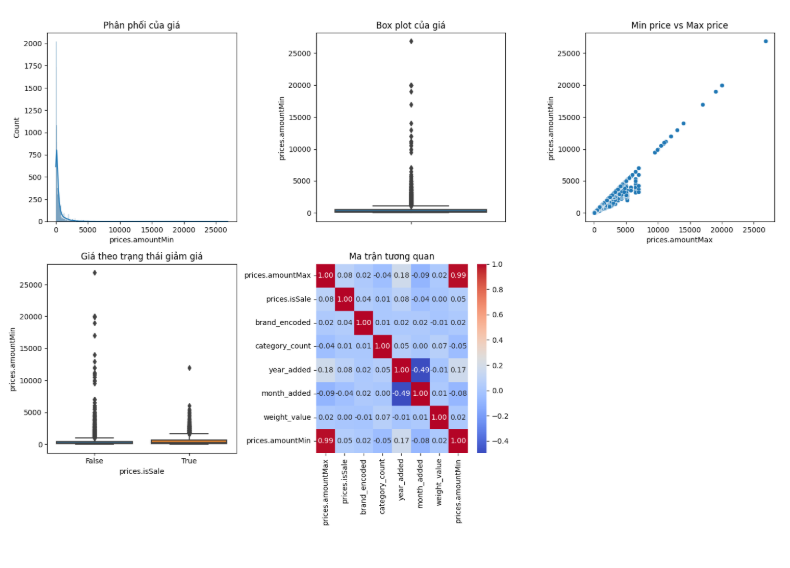
### 4.1.5. Mô phỏng kịch bản giá

Dashboard cho phép người dùng nhập các mức giá khác nhau và xem doanh thu dự kiến, hỗ trợ ra quyết định định giá linh hoạt.

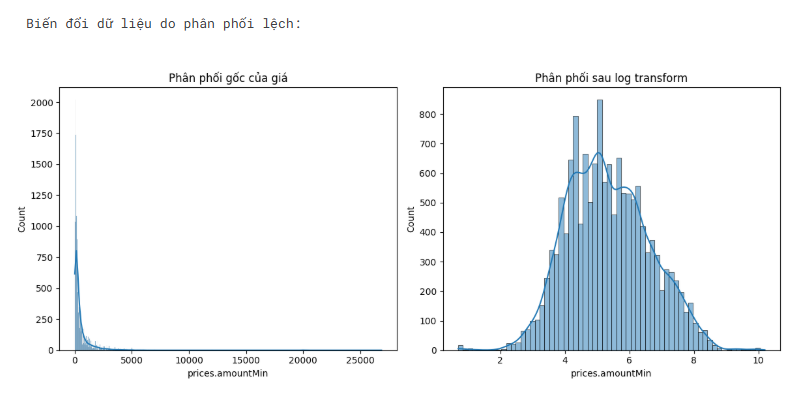
### 4.1.6. Trực quan hóa



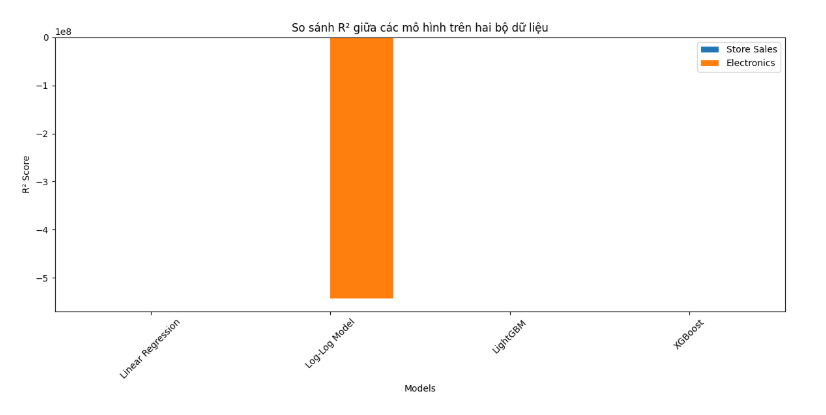
* **Phân phối lệch phải:** Đa số giao dịch có số lượng và giá trị bán thấp.
* **Khuyến mãi ảnh hưởng:** Giá trị bán hàng có xu hướng cao hơn khi có khuyến mãi.
* **Tương quan dương:** Số lượng bán tăng thì giá trị bán cũng có xu hướng tăng.
* **Tương quan yếu:** Khuyến mãi có tương quan dương yếu với giá trị bán



* Giá sản phẩm thường thấp, nhưng có một số ít sản phẩm giá rất cao.
* Giá khi không giảm thường cao hơn và biến động mạnh hơn giá khi giảm. Giá nhỏ nhất và lớn nhất của sản phẩm có xu hướng biến động cùng nhau.
* Việc giảm giá có xu hướng đi kèm với giá thấp hơn.



* Biểu đồ cho thấy phân phối giá gốc lệch mạnh về bên phải, tập trung ở các giá trị thấp. Sau khi log transform, phân phối trở nên đối xứng hơn, cho thấy log transform giúp giảm độ lệch và làm cho dữ liệu gần với phân phối chuẩn hơn.



* Mô hình tốt nhất cho Store Sales: XGBoost với R² = 0.8938
* Mô hình tốt nhất cho Electronics: XGBoost với R² = 0.9911
* **Kết luận:** Các mô hình hoạt động tốt hơn trên bộ dữ liệu Electronics
* **Phân tích chi tiết:**
  + 1. Các mô hình ensemble (LightGBM và XGBoost) nhìn chung hoạt động tốt hơn so với các mô hình tuyến tính truyền thống
  + 2. Log-Log transformation đã cải thiện hiệu suất của mô hình tuyến tính trên cả hai bộ dữ liệu

## 4.2 Các công nghệ sử dụng

* **Xử lý dữ liệu**:
  + pandas và numpy để thao tác và xử lý dữ liệu.
* **Xây dựng và đánh giá mô hình máy học**:
  + scikit-learn cho các mô hình cơ bản và công cụ đánh giá.
  + lightgbm và xgboost cho các mô hình ensemble hiệu suất cao.
* **Trực quan hóa dữ liệu**:
  + streamlit để xây dựng dashboard tương tác.
  + plotly, matplotlib, và seaborn để tạo biểu đồ trực quan.
* **Tối ưu hóa giá**:
  + Hàm minimize từ scipy.optimize với ràng buộc giá trong khoảng [min\_price, max\_price].

## 4.3 Xây dựng thành dashboard

* Dashboard được phát triển bằng streamlit, mang lại giao diện tương tác trực quan cho người dùng.
* Các chức năng chính:
  + Hiển thị phân tích dữ liệu và biểu đồ trực quan (phân phối giá, tương quan, ảnh hưởng khuyến mãi).
  + Mô phỏng kịch bản giá: Người dùng nhập mức giá và xem doanh thu dự kiến.
  + Hỗ trợ ra quyết định định giá linh hoạt và hiệu quả dựa trên kết quả mô hình và tối ưu hóa.

# Phần 5: Kết luận và đề xuất

## 5.1 Kết luận

Nghiên cứu về độ co giãn cầu dựa trên giá đã mang lại những hiểu biết sâu sắc về mối quan hệ giữa giá cả và hành vi tiêu dùng. Việc áp dụng đa dạng các mô hình, từ hồi quy tuyến tính, Log-Log đến các phương pháp học máy tiên tiến như Random Forest, LightGBM và XGBoost, đã giúp phân tích toàn diện các mối quan hệ tuyến tính và phi tuyến trong dữ liệu thị trường. Kết quả cho thấy sự khác biệt rõ rệt về độ co giãn giữa các ngành hàng, với một số ngành có tính co giãn cao (|PED| > 1) và một số khác ít co giãn (|PED| < 1), phản ánh đặc thù của từng thị trường.

Tiền xử lý dữ liệu, bao gồm xử lý giá trị thiếu và chuẩn hóa, đóng vai trò quan trọng trong việc nâng cao độ chính xác của các mô hình. Ngoài ra, nghiên cứu nhấn mạnh tầm quan trọng của các yếu tố không phải giá như thương hiệu, mùa vụ và khuyến mãi trong việc định hình phản ứng của người tiêu dùng. Mô hình tối ưu hóa giá dựa trên độ co giãn đã chứng minh tiềm năng cải thiện doanh thu thông qua chiến lược định giá thông minh, trong khi các chỉ số đánh giá (R², RMSE, MAE) và kỹ thuật Cross-Validation đảm bảo độ tin cậy của kết quả. Những phát hiện này không chỉ cung cấp cơ sở lý thuyết mà còn mang lại giá trị thực tiễn cho các chiến lược kinh doanh và hoạch định chính sách.

## 5.2 Đề xuất

Dựa trên kết quả nghiên cứu về độ co giãn cầu dựa trên giá, các đề xuất sau được đưa ra để nâng cao hiệu quả phân tích và ứng dụng thực tiễn. Thứ nhất, nên phát triển các mô hình phân tích động, như chuỗi thời gian hoặc mạng nơ-ron LSTM, để nắm bắt sự thay đổi của độ co giãn theo thời gian và dự đoán biến động trong tương lai. Thứ hai, mở rộng phạm vi nghiên cứu sang co giãn chéo và các yếu tố vĩ mô như lạm phát, đồng thời phân khúc khách hàng để cá nhân hóa phân tích co giãn theo hành vi mua sắm. Thứ ba, tích hợp kết quả vào chiến lược kinh doanh thông qua hệ thống định giá động, quản lý tồn kho và CRM để tối ưu hóa tiếp thị và bán hàng.

Ngoài ra, cần xây dựng các dashboard tương tác và công cụ mô phỏng trực quan để hỗ trợ ra quyết định. Về cải tiến mô hình, việc áp dụng ensemble learning, hyperparameter tuning tự động và giảm chiều dữ liệu sẽ nâng cao độ chính xác. Trong thực tiễn, thử nghiệm A/B testing, cơ chế phản hồi nhanh và hệ thống cảnh báo sớm sẽ giúp điều chỉnh chiến lược giá hiệu quả. Cuối cùng, đào tạo nhân viên và xây dựng hướng dẫn thực hành tốt nhất sẽ đảm bảo ứng dụng hiệu quả thông tin co giãn trong quyết định kinh doanh. Các đề xuất này sẽ góp phần nâng cao chất lượng phân tích và tối ưu hóa hiệu quả chiến lược kinh doanh.

# Tài liệu tham khảo

* Store Sales Price Elasticity Promotions Data, Kaggle.
* Datafiniti Electronics Products Pricing Data, Kaggle.
* Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research.