A white rectangular frame with black border

AI-generated content may be incorrect.

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**

A logo with a ball and a sphere

AI-generated content may be incorrect.

**BÁO CÁO**

**Đề tài: Dự đoán giá xe cũ ở thị trường Ấn Độ và các yếu tố ảnh hưởng**

Giảng viên hướng dẫn: ThS Lê Minh Tân

Sinh viên thực hiện

Nguyễn Hoàng Phúc 22110400

Tạ Nghĩa Nhân 22110388

Phạm Trung Kỳ 22110361

***Thành phố Hồ Chí Minh, tháng 03 năm 2025***

Mục Lục

[1. Lý do chọn đề tài 1](#_Toc193920947)

[2. Giới thiệu về Dataset 1](#_Toc193920948)

[3. Tiền xử lí dữ liệu và phân tích khám phá dữ liệu 4](#_Toc193920949)

[3.1. Tiền xử lí dữ liệu 4](#_Toc193920950)

[3.2. Khám phá dữ liệu 7](#_Toc193920951)

[4. Xây dựng mô hình 10](#_Toc193920952)

[5. Đánh giá mô hình 10](#_Toc193920953)

[6. Phân tích và đánh giá thực nghiệm 11](#_Toc193920954)

[7. Đánh giá mô hình 14](#_Toc193920955)

[8. Tài liệu tham khảo 15](#_Toc193920956)

# Lý do chọn đề tài

Thị trường xe cũ tại Ấn Độ đang phát triển mạnh mẽ trong những năm gần đây nhờ sự tăng trưởng kinh tế, dân số trẻ và sự bùng nổ của các nền tảng giao dịch trực tuyến. Sự chênh lệch giữa số lượng xe mới và nhu cầu mua xe với mức giá phải chăng đã tạo nên một thị trường xe cũ năng động, với số lượng giao dịch ngày càng tăng. Các nền tảng như CarDekho, OLX Autos và Cars24 đã góp phần hiện thực hóa sự chuyển đổi này, tạo điều kiện thuận lợi cho người mua và người bán.

Lý do chọn dự án này là vì việc hiểu rõ các đặc trưng ảnh hưởng đến giá xe cũ không chỉ giúp người tiêu dùng định giá hợp lý mà còn hỗ trợ các doanh nghiệp, nhà đầu tư và các bên liên quan trong ngành ô tô đưa ra các quyết định chiến lược. Phân tích dữ liệu để xác định các yếu tố chủ chốt như nhãn hiệu, năm sản xuất, số km đã chạy, loại nhiên liệu, dung tích động cơ,… sẽ giúp cải thiện độ chính xác trong dự đoán giá và tối ưu hóa quá trình giao dịch.

Dự án của chúng tôi tập trung vào việc khai thác dữ liệu xe cũ ở Ấn Độ, tiến hành tiền xử lý, trực quan hóa và áp dụng các mô hình học máy để xác định những yếu tố có ảnh hưởng lớn nhất đến giá bán xe. Mục tiêu của dự án là xây dựng một công cụ dự đoán giá chính xác, từ đó mang lại cái nhìn sâu sắc về thị trường xe cũ và hỗ trợ các quyết định giao dịch hiệu quả.

# Giới thiệu về Dataset

Dataset 1: Car Price Prediction Dataset

* Tác giả: Sukhmandeep Singh Brar
* Số lượng mẫu: 8128
* Số lượng đặc trưng: 12
* Nguồn dữ liệu: Kaggle, cardekho.com
* Loại kiểu dữ liệu: văn bản, số thực, số nguyên
* Tác dụng: Car Price Prediction Dataset bao gồm hàng loạt đặc điểm nổi bật liên quan đến ô tô, từ năm sản xuất, giá bán, số km đã lăn bánh, đến loại nhiên liệu, hình thức người bán, kiểu hộp số, số lượng chủ sở hữu trước đó, quãng đường đã đi, cùng các thông số kỹ thuật của động cơ. Những thuộc tính này mang đến cái nhìn sâu sắc và giá trị, hé lộ các yếu tố chi phối giá ô tô, đồng thời là nền tảng để xây dựng các mô hình dự đoán tinh tế, giúp ước lượng chính xác giá trị bán ra của những chiếc xe.
* Định dạng: CSV
* Danh sách các cột:
* name: tên, thương hiệu xe
* year: năm sản xuất (đơn vị: năm)
* selling\_price: giá bán (đơn vị: INR)
* km\_driven: số km đã đi
* fuel: loại nhiên liệu tiêu thụ
* seller\_type: loại người bán
* transmission: loại hộp số
* owner: số đời chủ sở hữu trước đó
* mileage: mức tiêu hao nguyên liệu của xe được đo bằng km/l hoặc km/kg
* engine: dung tích động cơ
* max\_power: công suất tối đa mà động cơ có thể tạo ra
* seats: số chỗ ngồi của xe
* Biến mục tiêu: selling\_price

Dataset 2: Used Car Prices

* Tác giả: Sujay R
* Số lượng mẫu: 5847
* Số lượng đặc trưng: 14
* Nguồn dữ liệu: Kaggle, Cars24
* Loại dữ liệu: văn bản, số thực, số nguyên
* Tác dụng: Used Car Prices chứa đựng nhiều chi tiết phong phú và các đặc điểm liên quan đến ô tô đã qua sử dụng. Trọng tâm của phân tích xoay quanh biến mục tiêu – giá của ô tô đã qua sử dụng, được tính bằng đơn vị lakhs, mang đến một cái nhìn sâu sắc về giá trị thị trường của những chiếc xe này.
* Định dạng: CSV
* Danh sách các cột
* Index: đánh thứ tự cho các hàng
* Name: tên, thương hiệu của xe
* Location: thành phố nơi chiếc xe được niêm yết
* Year: năm sản xuất
* Kilometers\_Driven: số km đã đi
* Fuel\_Type: loại nhiên liệu sử dụng
* Transmission: loại hộp số (tự động, thủ công)
* Owner\_Type: số đời chủ sở hữu trước đó
* Mileage: mức tiêu hao nguyên liệu của xe được đo bằng km/l hoặc km/kg
* Engine: dung tích động cơ
* Power: công suất tối đa
* Seats: số chỗ ngồi
* New\_Price: giá của chiếc xe mới cùng mẫu (đơn vị Cr hoặc Lakh)
* Price: giá bán hiện tại (đơn vị: Lakh)
* Biến mục tiêu: Price

Dataset 3: Car Price Prediction Dataset

* Tác giả: Jackson Divakar R (Owner)
* Số lượng mẫu: 2095
* Số lượng đặc trưng: 14
* Nguồn dữ liệu: Kaggle
* Loại dữ liệu: văn bản, số thực, số nguyên
* Tác dụng: Car Price Prediction Dataset được xây dựng nhằm ước tính giá của xe đã qua sử dụng dựa trên nhiều đặc điểm khác nhau. Dataset này hữu ích trong các bài toán máy học và phân tích hồi quy, giúp xác định các yếu tố tác động đến giá xe trên thị trường.
* Định dạng: CSV
* Danh sách các cột:
* #: đánh số thứ tự
* name: tên thương hiệu
* year: năm sản xuất
* selling\_price: giá bán (đơn vị INR)
* km\_driven: số km đã đi
* fuel: loại nhiên liệu sử dụng
* seller\_type: loại người bán
* transmission: loại hộp số (tự động hoặc thủ công)
* owner: số lượng người sở hữu trước đó
* seats: số chỗ ngồi
* max\_power: công suất tối đa
* Mileage Unit: đơn vị của Mileage
* Mileage: mức tiêu hao nguyên liệu của xe
* Engine: dung tích động cơ
* Biến mục tiêu: selling\_price

# Tiền xử lí dữ liệu và phân tích khám phá dữ liệu

## Tiền xử lí dữ liệu

Dữ liệu từ dataset 1:

* Loại bỏ các phương tiện sử dụng CNG và LPG bằng cách lọc những dòng không chứa hai loại nhiên liệu này.
* Loại bỏ xe thử nghiệm (Test Drive Car) bằng cách lọc bỏ những dòng có giá trị "Test Drive Car" trong cột owner.
* Xóa từ "Owner" trong cột owner bằng cách thay thế chuỗi " Owner" bằng rỗng.
* Lấy tên hãng xe bằng cách tách từ đầu tiên trong cột name.
* Chuyển đổi đơn vị giá bán bằng cách chia giá trị của cột selling\_price cho 100000.
* Đổi tên cột mileage (km/ltr/kg) thành mileage.
* Xóa cột seller\_type vì không cần thiết.

Dữ liệu từ dataset2:

* Loại bỏ xe điện bằng cách lọc bỏ những dòng có giá trị "Electric" trong cột Fuel\_Type.
* Trích xuất số từ các cột Engine, Power và Mileage bằng cách lấy giá trị số trong chuỗi dữ liệu.
* Loại bỏ các dòng có đơn vị "km/kg" trong cột Mileage.
* Lấy tên hãng xe bằng cách tách từ đầu tiên trong cột Name.
* Đổi tên các cột theo chuẩn, cụ thể:
  + Year đổi thành year
  + Kilometers\_Driven đổi thành km\_driven
  + Fuel\_Type đổi thành fuel
  + Transmission đổi thành transmission
  + Mileage đổi thành mileage
  + Engine đổi thành engine
  + Power đổi thành max\_power
  + Seats đổi thành seats
  + Price đổi thành selling\_price
  + Name đổi thành name
  + Owner\_Type đổi thành owner
* Xóa các cột không cần thiết gồm New\_Price, Location và Unnamed: 0.
* Sắp xếp lại thứ tự các cột theo danh sách chuẩn (correct\_order).

Dữ liệu từ dataset 3:

* Loại bỏ các phương tiện sử dụng CNG và LPG bằng cách lọc những dòng không chứa hai loại nhiên liệu này.
* Loại bỏ xe thử nghiệm (Test Drive Car) bằng cách lọc bỏ những dòng có giá trị "Test Drive Car" trong cột owner.
* Xóa từ "Owner" trong cột owner bằng cách thay thế chuỗi " Owner" bằng rỗng.
* Chuyển đổi đơn vị giá bán bằng cách chia giá trị của cột selling\_price cho 100000.
* Đổi tên các cột theo chuẩn, cụ thể:
  + max\_power (in bph) đổi thành max\_power
  + Mileage đổi thành mileage
  + Engine (CC) đổi thành engine
* Xóa các cột không cần thiết gồm seller\_type, Mileage Unit và Unnamed: 0.
* Sắp xếp lại thứ tự các cột theo danh sách chuẩn (correct\_order).

Dữ liệu từ dataset sau khi gộp

* Thay thế giá trị 0 trong cột mileage và max\_power thành NaN.
* Xử lý giá trị khuyết (NaN) của các cột kiểu số thực: Các giá trị NaN trong những cột này được thay thế bằng giá trị trung bình (mean) của từng cột.
* Xử lý giá trị khuyết của cột seats: Các giá trị NaN trong cột này được thay thế bằng giá trị mode (giá trị xuất hiện nhiều nhất).

## Khám phá dữ liệu

A group of graphs showing different types of data

AI-generated content may be incorrect.

Hình 3.2.1: Biểu đồ phân phối của một số feature

A graph with different colored squares

AI-generated content may be incorrect.

Hình 3.2.2: Biểu đồ skewness của một số feature

A blue and white grid with black text

AI-generated content may be incorrect.

Hình 3.2.3: Ma trận quan hệ giữa các feature

A graph with blue squares

AI-generated content may be incorrect.

Hình 3.2.4: Biểu đồ thể hiện mức độ quan hệ với Selling Price với các feature khác

A group of pie charts

AI-generated content may be incorrect.

Hình 3.2.5: Biểu đồ tròn thể hiện tỉ trọng của các biến trong các feature

# Xây dựng mô hình

Danh sách các mô hỉnh được chọn để huấn luyện

* Linear Regression
* Ridge Regression
* Lasso Regression
* Decision Tree
* K-Nearest Neighbors (KNN)
* Random Forest
* Support Vector Machine (SVM)

Lý do chọn những mô hỉnh trên

* Linear Regression được sử dụng để đánh giá tính tuyến tính của dữ liệu.
* Nếu xuất hiện hiện tượng overfitting, Ridge Regression hoặc Lasso Regression sẽ được áp dụng để điều chỉnh.
* Trong trường hợp dữ liệu phức tạp và có tính phi tuyến tính, các mô hình như Decision Tree, KNN hoặc Random Forest sẽ được xem xét.
* Random Forest thường đạt độ chính xác cao nhất, nhưng cần cân nhắc tài nguyên tính toán.

Mô hình tối ưu sẽ được chọn dựa trên chỉ số Test Mean Squared Error (MSE) thấp nhất, đảm bảo khả năng dự đoán tốt nhất trên tập dữ liệu kiểm tra.

# Đánh giá mô hình

Các mô hình sẽ được đánh giá dựa trên chỉ số MSE:

- Mean Absolute Error (MSE): Đo lường giá trị trung bình của sai số tuyệt đối giữa giá trị thực tế và giá trị dự đoán. Giá trị MSE càng nhỏ độ chính xác của mô hình càng cao.

A screenshot of a graph

AI-generated content may be incorrect.

Hình 5.1: Bảng giá trị thể hiện MSE của từng mô hình

# A graph of different colored bars AI-generated content may be incorrect.Phân tích và đánh giá thực nghiệm

Hình 6.1: Biểu đồ so sánh chỉ số MSE của các mô hình

A graph with blue dots and red line

AI-generated content may be incorrect.

Hình 6.2: Biểu đồ phân tán thể hiện mối quan hệ giữa giá thực tế và giá dự đoán của mô hình Random Forest

A graph of different colored lines

AI-generated content may be incorrect.

Hình 6.3: Biểu đồ phân tán thể hiện mối quan hệ giữa giá thực tế và giá dự đoán của mô hình Random Forest

Từ Hình 6.3, ta có thể dễ dàng nhận thấy model đạt hiệu suất cao nhất dựa vào chỉ số MSE. Đó là Random Forest.

Với lí thuyết cung cấp ở phần Xây dựng mô hình, ta có thể đoán được Random Forest là sẽ là mô hình hiệu quả nhất cho bài toán của chúng ta đã đề ra. Bằng thực nghiệm Hình 6.2 và Hình 6.3 thì điều đó càng được khẳng định hơn.

Random Forest sẽ là mô hình tốt nhất cho bài toán của chúng ta. Từ đó, ta sẽ đánh giá mức độ ảnh hưởng của các đặc trưng quan trọng đối với giá.

A graph with blue bars

AI-generated content may be incorrect.

Hình 6.4: Biểu đồ thể hiện mức độ ảnh hưởng đến giá của các đặc trưng

Với Hình 6.4, max\_power sẽ ảnh hưởng nhiều nhất đến giá, tiếp đến là year, km\_driven,… và cuối cùng là owner.

# Đánh giá mô hình

Thông qua quá trình phân tích dữ liệu và xây dựng mô hình, chúng tôi phát hiện rằng **công suất cực đại** chính là yếu tố quan trọng nhất quyết định giá xe ô tô. Nói một cách đơn giản, những chiếc xe sở hữu động cơ mạnh mẽ hơn thường có giá cao hơn trên thị trường. Dù các yếu tố khác như năm sản xuất (tuổi xe), số km đã đi hay dung tích động cơ cũng góp phần tạo nên sự khác biệt về giá, nhưng không gì vượt qua được tầm ảnh hưởng của **công suất cực đại**.

**Vì sao công suất cực đại lại “lên ngôi”?**

Lý do rất rõ ràng: **công suất cực đại** phản ánh trực tiếp sức mạnh và hiệu suất của xe. Đây là yếu tố quyết định khả năng tăng tốc, tốc độ tối đa và cảm giác lái phấn khích mà người dùng mong đợi. Người mua xe thường sẵn sàng chi trả nhiều hơn để sở hữu những chiếc xe mạnh mẽ, và điều này đẩy giá trị thị trường của các mẫu xe có công suất cao lên đáng kể.

**Các yếu tố khác đóng vai trò thế nào?**

Dù **công suất cực đại** dẫn đầu, những yếu tố khác vẫn có tiếng nói riêng trong việc định hình giá xe:

* **Năm sản xuất (tuổi xe):** Xe càng cũ, giá càng giảm do khấu hao – điều này không có gì bất ngờ. Tuy nhiên, tác động của nó thường mờ nhạt hơn so với sức hút từ hiệu suất.
* **Số km đã đi:** Quãng đường dài đồng nghĩa với độ hao mòn lớn, khiến giá xe giảm. Nhưng ngay cả vậy, nó vẫn không đủ sức “soán ngôi” tầm quan trọng của công suất.
* **Dung tích động cơ:** Dù liên quan mật thiết đến công suất, dung tích động cơ lại không thể hiện đầy đủ sức mạnh thực tế như **công suất cực đại**.

**Lợi ích thực tế từ phát hiện này**

Hiểu được vai trò của **công suất cực đại** mang lại lợi thế lớn cho tất cả các bên:

* **Người mua và người bán:** Người bán có thể tự tin định giá cao hơn nếu xe có công suất vượt trội, trong khi người mua sẽ dễ dàng đánh giá xem mức giá có xứng đáng với hiệu suất hay không.
* **Doanh nghiệp:** Các đại lý hoặc nhà sản xuất có thể dựa vào xu hướng này để tối ưu hóa chiến lược định giá, tập trung quảng bá những mẫu xe mạnh mẽ để thu hút khách hàng và tăng lợi nhuận.

**Hướng đi để cải thiện**

Để dự đoán giá xe chính xác hơn, chúng ta có thể bổ sung thêm các yếu tố như thương hiệu, loại nhiên liệu hay kiểu hộp số. Đồng thời, việc ứng dụng các mô hình học máy tiên tiến hơn sẽ giúp khám phá những xu hướng ẩn sâu, mang lại kết quả sắc nét và đáng tin cậy hơn.

Tóm lại, **công suất cực đại** không chỉ là “vua” trong việc định giá xe mà còn phản ánh rõ nét sở thích của thị trường đối với hiệu suất. Dựa trên phát hiện này, người mua, người bán và doanh nghiệp có thể tự tin bước đi trên thị trường xe hơi, trong khi những cải tiến trong tương lai hứa hẹn sẽ mang đến sự chính xác và hiệu quả cao hơn nữa.

# Tài liệu tham khảo

Learning Data Science **By**[**Sam Lau**](http://www.samlau.me/)**,**[**Joey Gonzalez**](https://people.eecs.berkeley.edu/~jegonzal/)**, and**[**Deb Nolan**](https://www.stat.berkeley.edu/~nolan/)**.**