**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

🙤🙧🟍🙥🙦



**Phân tích và dự đoán dữ liệu giá xe ô tô đã qua sử dụng bằng Random Forest**

|  |  |  |  |
| --- | --- | --- | --- |
| **Nhóm 8** | | | |
| Sinh viên thực hiện: | | | |
| STT | Họ tên | MSSV | Ngành |
| 1 | Nguyễn Việt Quang | 21522515 | CNTT |
| 2 | Trịnh Tuấn Tú | 21522747 | CNTT |

**TP. HỒ CHÍ MINH – 12/2024**

# GIỚI THIỆU

Đề tài trong đồ án này nhằm phân tích thăm dò các đặc trưng ảnh hưởng đến giá xe và sau đó xây dựng mô hình máy học random forest dự đoán xe ô tô có độ chính xác cao.

Đề tài này được thực hiện bằng cách sử dụng thư viện Pandas, Numpy để xử lý dữ liệu thông qua các phương pháp như tiền xử lý dữ liệu gồm xử lý dữ liệu bị khuyết, xử lý các giá trị outliers; phân tích thăm dò sự tương quan của các đặc trưng; Dùng thư viện sklearn để xây dựng mô hình học máy Random Forest; đánh giá hiệu suất bằng việc sử dụng độ đo R2 score, MSE (Mean Squared Error) để đánh giá mô hình dự đoán so với thực tế.

Kết quả đạt được là xây dựng một mô hình với độ chính xác trên tập test cao. Từ đó có thể thấy rằng phương pháp phân tích thăm dò và xử lý bộ dữ liệu đạt kết quả tốt.

Trong đề tài này, nhóm đã sử dụng bộ dữ liệu [Used Cars Price Prediction | Kaggle](https://www.kaggle.com/datasets/avikasliwal/used-cars-price-prediction?select=train-data.csv). Chúng tôi đã tự tìm kiếm bộ dữ liệu trên Kaggle [1] và đề tài được hoàn thành không phụ thuộc vào bất kì nguồn bên ngoài.

# MÔ TẢ BỘ DỮ LỆU

## Mô tả:

Bộ dữ liệu xe ô tô cũ là của tác giả Avi Kasliwal, nhóm đã tham khảo tại [link dataset](https://www.kaggle.com/datasets/avikasliwal/used-cars-price-prediction?select=train-data.csv) và chọn bộ dữ liệu này để thực hiện đồ án. Bộ dữ liệu có tên là “Used Cars Price Prediction” được tác giả thu thập và update lần cuối là vào 5 năm trước. Mục đích tác giả xây dựng bộ dữ liệu nhằm dự đoán giá của một chiếc xe chưa biết.

Bộ dữ liệu xe ô tô cũ chứa thông tin về những chiếc ô tô cũ đã qua sử dụng bao gồm các thuộc tính như tên thương hiệu, kiểu xe, vị trí được rao bán, phiên bản, loại nhiên liệu xe sử dụng, số lượng ghế trên xe, ... Bộ dữ liệu gồm có 14 cột, 6019 dòng, có 5 biến phân loại và 9 biến số, kích thước 662.72 kB. Dưới đây là mô tả chi tiết các thuộc tính:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **STT** | **Thuộc tính** | **Kiểu dữ liệu** | **Loại biến** | **Ý nghĩa** | **Khoảng giá trị** |
| 1 | Unnamed | int | 1 | STT | 0 🡪 6018 |
| 2 | Name | object | 2 | Thương hiệu và kiểu xe. | 1876 giá trị khác nhau |
| 3 | Location | object | 2 | Địa điểm mà xe đang được bán hoặc sẵn sàng để mua. | Mumbai 13%,  Hyderabad 12%, Other 75% |
| 4 | Year | int | 1 | Năm hoặc phiên bản của mẫu xe. | 1998 🡪 2019 |
| 5 | Kilometers\_Driven | int | 1 | Tổng số km đã đi của chủ sở hữu trước tính bằng KM. | 171 🡪 6500000 |
| 6 | Fuel\_Type | object | 2 | Loại nhiên liệu mà xe sử dụng. (Xăng/Dầu diesel/Điện/CNG/ LPG) | Diesel 53%,  Petrol 46%,  Other 1% |
| 7 | Transmission | object | 2 | Loại hộp số mà xe sử dụng. (Tự động/Số tay) | Manual 71%  Automatic 29% |
| 8 | Owner\_Type | object | 2 | Chủ sở hữu là xe cũ, xe đã qua sử dụng hay loại khác. | First 82%  Second 16%  Other 2% |
| 9 | Mileage | object | 1 | Số km tiêu chuẩn mà hãng xe cung cấp tính bằng kmpl hoặc km/kg | 18.9 kmpl 3%  17.0 kmpl 3%  Other 94% |
| 10 | Engine | object | 1 | Dung tích xi lanh của động cơ tính bằng cc. | 1197 CC 10%  1248 CC 9%  Other 81% |
| 11 | Power | object | 1 | Công suất tối đa của động cơ tính bằng mã lực. | 74 bhp 4%  98.6 bhp 2%  Other 94% |
| 12 | Seats | float | 1 | Số ghế trên xe. | 0 🡪 10 |
| 13 | New\_Price | object | 1 | Giá của một chiếc xe mới cùng mẫu. | [null] 86%  Other 14% |
| 14 | Price | float | 1 | Giá của một chiếc xe đã qua sử dụng tính bằng INR Lakhs. | 0.44 🡪 160 INR Lakhs. |

***\* Ghi chú:***

***1 – Biến số***

***2 – Biến phân loại***

Thống kê số lượng khuyết của bộ dữ liệu:

|  |  |
| --- | --- |
| **Tên cột** | **Số lượng khuyết** |
| Unnamed | 0 |
| Name | 0 |
| Location | 0 |
| Year | 0 |
| Kilometers\_Driven | 0 |
| Fuel\_Type | 0 |
| Transmission | 0 |
| Owner\_Type | 0 |
| Mileage | 2 |
| Engine | 36 |
| Power | 36 |
| Seats | 42 |
| New\_Price | 5195 |
| Price | 0 |

# PHƯƠNG PHÁP PHÂN TÍCH

Quy trình phân tích dữ liệu:



Hình 1: Tổng quan quy trình phân tích dữ liệu

## Tìm kiếm dữ liệu:

Sau khi tìm kiếm kỹ lưỡng, nhóm đã phát hiện ra một tập dữ liệu để dự đoán giá xe dựa trên nhiều tính năng khác nhau có tên Used Car Prediction

## Xác định vấn đề:

Trong thị trường xe hơi thay đổi nhanh chóng như hiện nay, việc dự đoán giá xe rất quan trọng đối với cả người mua và người bán. Dự án này nhằm mục đích phân tích giá xe ô tô đã qua sử dụng và có thể tạo ra một mô hình học máy để dự đoán giá xe dựa trên nhiều tính năng khác nhau.

## Tiền xử lý bộ dữ liệu:

Sau khi xem tổng quan bộ dữ liệu, thực hiện xóa biến không cần thiết “Unamed: 0”. Các biến “Mileage”, “Power”, “Engine” chứa dữ liệu bao gồm cả số và đơn vị đi kèm, ngoài ra kiểu dữ liệu cũng không đúng. Vì thế, nhóm đã thực hiện xóa bỏ các đơn vị đi kèm và đổi kiểu dữ liệu sang thành “Float”.

Thực hiện kiểm tra số lượng dữ liệu khuyết thì phát hiện dữ liệu khuyết ở các biến “Mileage”, “Engine”, “Power”, “Seats”, “New\_Price”. Ở biến “New\_Price” thì số lượng khuyết lên đến hơn 86% và cũng không cần thiết nên thực hiện xóa biến “New\_Price”. Các biến còn lại, số lượng dữ liệu khuyết khá ít nên tiến hành điền khuyết.

Dựa vào các hình ảnh trực quan dữ liệu theo từng biến, nhóm đã chọn các phương pháp điền khuyết khác nhau. Ở biến “Mileage” do dữ liệu có phân phối gần chuẩn, có thể thay thế giá trị khuyết bằng mean của cột. Hai biến “Engine” và “Power” do phân phối dữ liệu lệch phải nên có thể thay thế giá trị khuyết bằng median. Biến “Seats” thì sử dụng giá trị xuất hiện nhiều nhất (mode) để điền khuyết vì đây thường là số ghế phổ biến nhất.

Nhận thấy ở biến “Name” dữ liệu bao gồm tên thương hiệu và model nên tiến hành tạo biến “Brand” và “Model” mới từ biến “Name”, sau đó xóa biến “Name”. Tạo thêm biến “Age” chứa dữ liệu về độ tuổi của xe.

## Phân tích thăm dò dữ liệu

Trong quá trình này, nhóm đã thực hiện các bước cơ bản như xem tổng quan dữ liệu sau khi tiền xử lý, kiểm tra lại dữ liệu khuyết và kiểu dữ liệu của từng biến. Tính toán các phân vị (quantiles) cho tất cả các cột dạng số. Phát hiện các biến có giá trị outliers nên tiến hành loại bỏ.

Sau đó, thực hiện phân tích các biến số và biến phân loại trong bộ dữ liệu biến nào tương đối gần mức đối xứng, các biến nào đang mất cân đối. Trực quan hóa bằng biểu đồ cột để xem phân phối dữ liệu của từng biến. Phát hiện các biến “Kilometers\_Driven”, “Engine”, “Power”, “Price” có dấu hiệu mất cân đối, nhóm đã sử dụng thêm Skewness để kiểm tra xem các biến có mức độ lệch thế nào. Kết quả kiểm tra Skewness của các biến “Kilometers\_Driven”, “Engine”, “Power”, “Price” cho thấy tất cả đều lệch phải. Xử lý tất cả các biến lệch phải, dùng Log Transformation để xử lý biến “Price”, dùng Square Root Transformation để xử lý biến “Power” và “Engine”, sử dụng Box-Cox Transformation để xử lý biến Kilometers\_Driven.

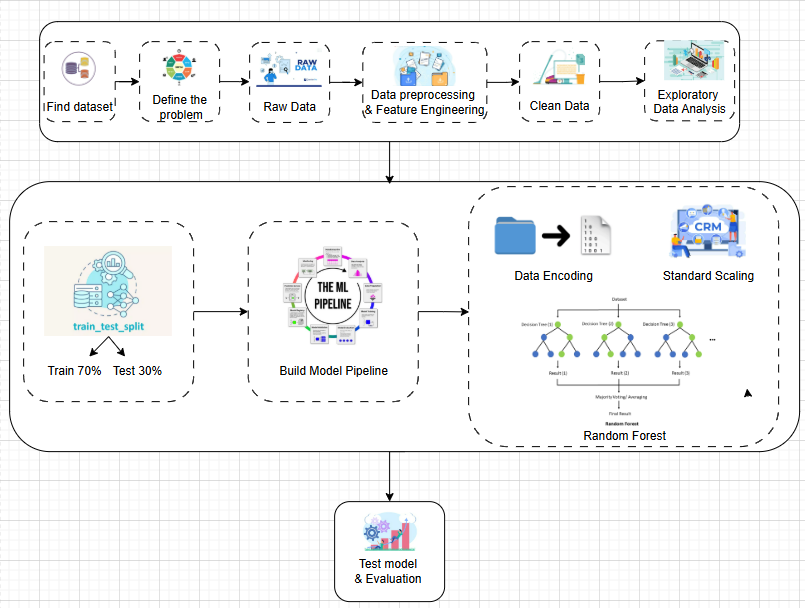
Phân tích thăm dò tìm ra các biến có khả năng ảnh hưởng đến giá xe. Các biến số và phân loại quan trọng nào ảnh hưởng đến giá xe. Sử dụng biểu đồ nhiệt và cột để trực quan hóa dữ liệu các biến số với biến “Price” để xem mối tương quan. Từ đó rút ra được các biến số quan trọng nào ảnh hưởng đến giá xe. Sử dụng p-value để tìm các biến phân loại ảnh hưởng đến giá xe. Biến số và biến phân loại nào ảnh hưởng nhất đến giá xe. Cuối cùng, trực quan hóa bằng Power BI, Seaborn, Matplotlib

## Mô hình hóa dữ liệu

Trước hết cần mã hóa các dữ liệu không phải kiểu số bằng phương pháp One-Hot Encoding và sử dụng Standard Scaling cho các biến số để chuẩn hóa chúng, đảm bảo rằng chúng có phạm vi giá trị đồng nhất, giúp mô hình học tốt hơn. Tiếp theo, chia tập dữ liệu thành tập train chiếm 70% và test chiếm 30%. Sau đó, xây dựng mô hình học máy Random Forest Regressor. Cuối cùng, tạo pipeline kết hợp tất cả các bước trên thành một pipeline duy nhất.

## Đánh giá hiệu suất

Sử dụng mô hình tốt nhất sau khi huấn luyện trên tập train. Dự đoán kết quả trên tập test. Đánh giá kết quả dự đoán so với thực tế bằng độ đo MSE, R2 score



Hình 2: Chi tiết thực hiện toàn bộ quy trình

# PHÂN TÍCH THĂM DÒ

Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, Sơ đồ

Mô tả được tạo tự động

Hình 3: Dashboard dữ liệu

Các kết quả phân tích thăm dò cho thấy bộ dữ liệu bao gồm 30 thương hiệu xe khác nhau với tổng số 6,019 chiếc, trong đó trung bình mỗi chiếc xe có khoảng 4 đến 6 chỗ ngồi. Giá trị trung bình của mỗi chiếc xe là 9.48K INR Lakhs, và tổng giá trị lên đến 57 triệu INR Lakhs. Nguồn nhiên liệu được ưa chuộng là Diesel và CNG.

Bộ dữ liệu đã được làm sạch và không còn giá trị khuyết. Tuy nhiên, các biến như Price, Kilometers\_Driven, Engine và Power có sự mất cân bằng với phân phối lệch phải. Trong số các biến phân loại, Transmission là yếu tố ảnh hưởng mạnh nhất đến giá xe, khi các xe có Transmission là Automatic luôn có giá cao hơn so với Manual.

Xét các biến số, những yếu tố ảnh hưởng đến giá xe bao gồm Power, Engine, Age và Mileage, trong đó Power là yếu tố có ảnh hưởng lớn nhất và thể hiện mối quan hệ tỉ lệ thuận với Price. Ngược lại, Kilometers\_Driven có ảnh hưởng ít nhất đến giá. Đối với các biến phân loại, Owner\_Type là yếu tố ít ảnh hưởng nhất tới giá xe.

# KẾT QUẢ PHÂN TÍCH

Kết quả phân tích thăm dò ảnh hưởng của các biến tới giá xe sắp xếp theo thứ tự từ cao đến thấp

|  |  |
| --- | --- |
| **Biến số** | **Biến phân loại** |
| Power  Engine  Age  Year  Mileage  Seats  Kilometers\_Driven | Transmission  Brand  Model  Fuel\_Type  Location  Owner\_Type |

Sau khi xem xét các giá trị outliers, nhóm đã thực hiện xử lý các giá trị outliers như loại bỏ các giá trị vượt quá phân vị 99% ở biến “Kilometers\_Driven”, “Price” và “Power”. Giới hạn biến “Age” trong khoảng 20 năm, loại bỏ số ghế là 0 ở biến “Seats”. Phát hiện và xử lý các biến “Power”, “Engine”, “Price”, “Kilometers\_Driven” bị lệch phải.

Xét về ảnh hưởng của các biến phân loại, Transmission là yếu tố có tác động mạnh nhất đến giá xe, khi các xe có Transmission là Automatic luôn có giá cao hơn so với Manual. Điều này có thể xuất phát từ sự tiện lợi và công nghệ cao cấp hơn của xe số tự động. Trong khi đó, Owner\_Type là biến phân loại có ảnh hưởng ít nhất đến giá xe.

Đối với các biến số, Power, Engine, Age và Mileage là những yếu tố quan trọng ảnh hưởng đến giá xe. Đặc biệt, Power có mối quan hệ tỉ lệ thuận rõ ràng với Price, cho thấy những chiếc xe có công suất cao thường có giá trị cao hơn. Ngược lại, Kilometers\_Driven có ảnh hưởng ít nhất đến giá, điều này có thể giải thích bởi giá trị xe không giảm đáng kể dựa trên quãng đường đã di chuyển nếu các yếu tố khác như động cơ và công suất vẫn còn tốt.

Ngoài ra, xu hướng giá theo thời gian cũng cho thấy sự gia tăng đáng kể trong những năm gần đây, đặc biệt ở các dòng xe đời mới với công nghệ tiên tiến hơn. Giá xe có xu hướng cao hơn đối với các dòng xe trẻ tuổi và có công suất mạnh, trong khi những chiếc xe cũ hơn với quãng đường đã di chuyển lớn có mức giá thấp hơn.

Tổng thể, kết quả phân tích chỉ ra rằng các yếu tố công nghệ và hiệu suất như Transmission và Power đóng vai trò quyết định trong việc định giá xe, trong khi các yếu tố phân loại khác ít ảnh hưởng hơn.

# KẾT LUẬN

Dự án dự đoán giá xe ô tô đã qua sử dụng được thực hiện qua các bước từ thu thập, tiền xử lý dữ liệu, đến phân tích và mô hình hóa. Dữ liệu được làm sạch bằng cách loại bỏ các biến không cần thiết, xử lý dữ liệu khuyết bằng các phương pháp mean, median, mode phù hợp với phân phối dữ liệu, và mã hóa dữ liệu không phải số. Các biến lệch phải như “Price”, “Power”, “Engine”, và “Kilometers\_Driven” được điều chỉnh bằng các phép biến đổi như Log, Square Root, và Box-Cox. Phân tích thăm dò xác định các biến quan trọng ảnh hưởng đến giá xe thông qua kiểm tra tương quan và p-value.

Mô hình Random Forest Regressor được xây dựng với dữ liệu chuẩn hóa và chia tỷ lệ train-test 70-30. Hiệu suất mô hình được đánh giá qua MSE và R2-score, cho thấy khả năng dự đoán giá xe hiệu quả. Kết quả Mean Squared Error: 0.05, R2 Score: 0.90.

Trực quan hóa dữ liệu bằng Power BI và Seaborn đã hỗ trợ tìm hiểu rõ ràng mối quan hệ giữa các biến.

**TÀI LIỆU THAM KHẢO**

Kaggle. Link: [Dataset-Used-Cars-Price](https://www.kaggle.com/datasets/avikasliwal/used-cars-price-prediction?select=train-data.csv) (22/10/2024)

Power BI. Link: <https://www.microsoft.com/en-us/power-platform/products/power-bi> (10/12/2024)

**PHỤ LỤC PHÂN CÔNG NHIỆM VỤ**

|  |  |  |
| --- | --- | --- |
| **STT** | **Thành viên** | **Nhiệm vụ** |
| 1 | Nguyễn Việt Quang | Thực hiện toàn bộ mục 1,4,5,6 |
| 2 | Trịnh Tuấn Tú | Thực hiện toàn bộ mục 2,3. Chỉnh sửa thêm phần 5 và 6. |