ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

NGÀNH: KHOA HỌC DỮ LIỆU

MÔN HỌC: LẬP TRÌNH CHO KHOA HỌC DỮ LIỆU

MÔ HÌNH DỰ ĐOÁN GIÁ XE MÁY

Sinh viên:
Nguyễn Duy Anh – 20002028
Lê Thế Cường – 20002035
Vũ Trọng Quân – 20002087

Giảng viên hướng dẫn:
Ngô Thế Quyền

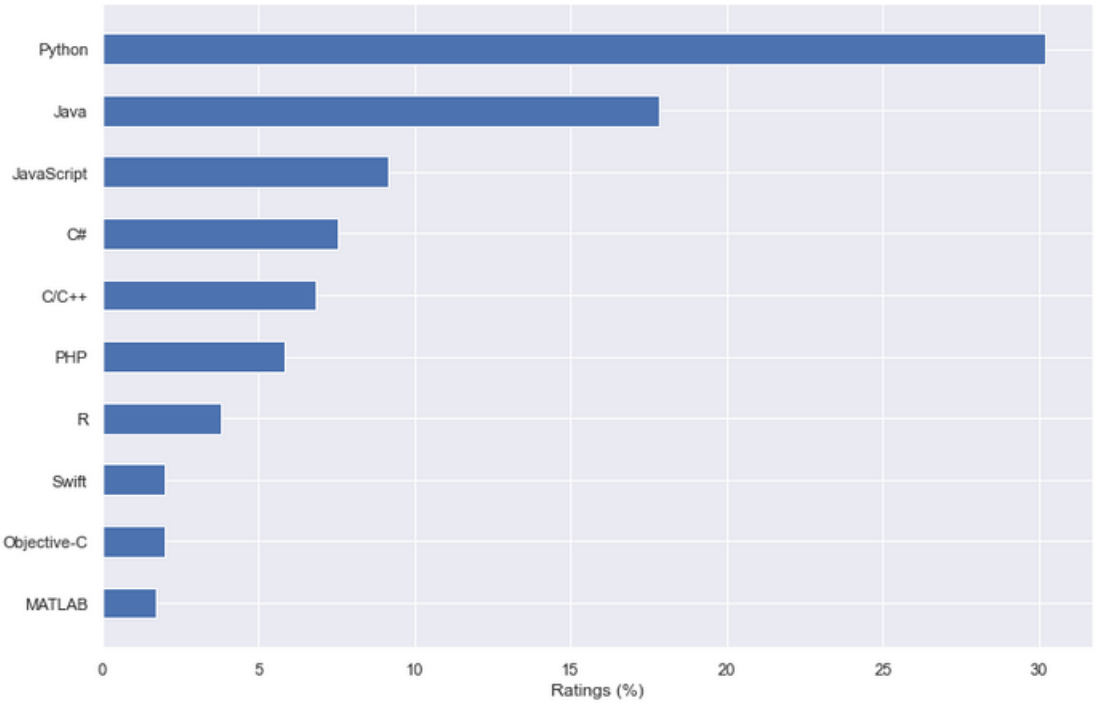
Lời nói đầu

Khoa học dữ liệu là ngành khoa học sử dụng các phương pháp, thuật toán để tìm hiểu về các đặc trưng của dữ liệu, phân tích ra ý nghĩa ẩn sâu bên trong chúng, từ đó đưa ra những quyết định, mang lại những lợi ích cho cá nhân hoặc tổ chức. Trong những năm gần đây, lợi ích của khoa học dữ liệu tạo ra ngày càng lớn. Điều này đến từ một phần không nhỏ trong sự phát triển mạnh mẽ của các công cụ tính toán trên các phần mềm máy tính. Ngôn ngữ lập trình Python là một trong những công cụ như vậy.

Python là một ngôn ngữ lập trình, hoạt động theo một nguyên lý và có các cấu trúc lập trình giống như bao ngôn ngữ khác như C++, Java, JavaScript. Nhưng đối với những người trong lĩnh vực Khoa học dữ liệu và Trí tuệ nhân tạo nói chung, nó là một ngôn ngữ đặc biệt. Sự tinh tế, đơn giản trong cú pháp, câu lệnh, kết hợp với phong phú của các thư viện (nhất là các thư viện dùng cho phân tích dữ liệu, tính toán ma trận học máy, học sâu...) làm cho Python trở thành một "người bạn đồng hành"đúng nghĩa đối với "dân trong nghề này".

Trong bài báo cáo này về dự án lần này, một dự án về hệ thống học máy, xây dựng nên một phần mềm với giao diện là một trang web có tích hợp một mô hình học máy để dự đoán ra giá của một chiếc xe máy khi biết các đặc điểm của nó, nhóm chúng tôi sẽ sử dụng Python để thực hiện nó một cách trọn vẹn, trong tất cả các khâu xử lý, từ việc chuẩn bị nguồn dữ liệu, cho đến khi tạo ra sản phẩm, để chúng ta thấy được rằng: Trong Khoa học dữ liệu hiện đại, Python hữu ích đến nhường nào. Và nếu phải chọn duy nhất một ngôn ngữ để thực hiện hoàn chỉnh một dự án Khoa học dữ liệu, chắc chắc Python là công cụ hiếm hoi có thể thực hiện điều đó (tất nhiên có thể không phải là duy nhất).

Chính vì vậy, khi một người muốn bước chân vào con đường Khoa học dữ liệu, nghiên cứu cách phân tích số liệu, hay nghiên cứu các mô hình học máy, học sâu, tôi khuyên rằng họ không nên bỏ qua sự tiếp cận với Python. Cộng đồng của ngôn ngữ này rất đông đảo, nên bạn cũng sẽ có thể dễ dàng tìm được những hướng giải quyết nhờ những sự chia sẻ của mọi người trong cộng đồng thông qua những trang web trên mạng (Ví dụ như StackOverFlow,...). Theo đánh giá của trang DataCamp, một trang chuyên dụng đào tạo kỹ năng khoa học dữ liệu(https://www.datacamp.com/blog/top-programming-languages-for-data -scientists-in-2022) Python là ngôn ngữ có tỉ lệ yêu thích cao nhất.



Hình 1: Tỷ lệ yêu sử dụng ngôn ngữ lập trình Python

Rất may mắn rằng, khi được đào tạo các lĩnh vực chuyên ngành dành cho Khoa học dữ liệu, nhóm chúng tôi đã may mắn được tiếp cận với Python thông qua môn học "Lập trình cho Khoa học dữ liệu", Nhóm 4 chúng em xin được cảm ơn thầy Ngô Thế Quyền, giảng viên phụ trách hướng dẫn của môn học. Bài bào cáo này sẽ là một bản trình bày, thể hiện được những gì cả nhóm đã học được, đã đạt được trong quá trình học, tất cả đều được thể hiện thông qua dự án.

Mục lục

1 Phát biểu vấn đề và đưa ra ý tưởng dự án giải quyết

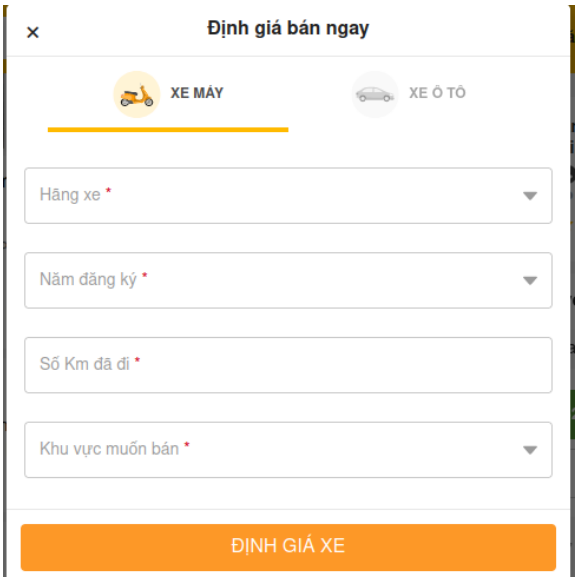
Trong thời đại số, việc mua bán online trên các trang thương mại điện tử, hay các trang web chuyên dụng về một sản phẩm diễn ra ngày càng phổ biến. Các trang web này cho phép người dùng có thể truy cập, tạo một tài khoản cá nhân, kết nối với các tài khoản thẻ ngân hàng hay thẻ tín dụng để mua đồ và cũng cho phép mọi người đăng lên các món đồ muốn bán.

Mua bán lại xe máy, đặc biệt là xe đã qua sử dụng một thời gian, là một mặt hàng có sự trao đổi, mua bán diễn ra hết sức thường xuyên trong các hoạt động mua bán online, việc phổ biến của việc mua bán xe máy thúc đẩy sự ra đời của các nền tảng, các trang thương mại điện tử chỉ chuyên dụng cho việc bán xe, một số rất nổi tiếng có thể kể đến như trang Okxe, Chợ tốt xe máy, Xemaycugiare, muaban.net (Người đọc có thể tra google với những cái tên này).

Tuy nhiên, khi mua - bán xe máy cũng có một vấn đề phát sinh: *Làm sao để có thể biết được giá bán hợp lý khi ta đăng bán một chiếc xe cũ? Liệu cái giá ta đưa ra có quá rẻ, nếu trường hợp như vậy thì người bán sẽ có phần thiệt thòi, còn liệu với những cái giá quá cao, chắc chắn, sản phẩm sẽ khó bán hơn. Nhưng khi bán được, ngược lại, điều thiệt thòi lại đổ về người mua hàng. Không chỉ có người bán, người mua cũng có nhu cầu muốn biết về một mức giá hợp lý khi mua lại xe, để tránh sự thiệt thòi dành về phía mình.* Đây cũng chính là một vấn đề đã thúc đẩy nhóm 4 thực hiện, tạo ra dự án xây dựng một hệ thống website ứng dụng một mô hình học máy giúp dự đoán giá xe máy dựa vào các thông số của xe để giải quyết vấn đề một cách thỏa đáng.

Hiểu được những vấn đề mà người bán, người mua đều gặp phải, các trang thương mại điện tử nói chung và các trang chuyên dụng về xe máy nói chung đang cố gắng nghiên cứu, tìm ra một hướng đi tốt nhất cho cả hai bên. Phương pháp chủ yếu được họ tính đến cũng là tích hợp vào trang web một hệ thống học máy, để đưa ra mức giá phù hợp dựa vào các đặc điểm như hãng xe, số km đã đi, dung tích xe,... Ví dụ như đối với trang xe Chợ tốt, trên trang web của xe có một chức năng, tiêu đề "dịch vụ và tiện ích", trong đó có một mục Định giá xe cũ hỗ trợ cho người mua và người bán, muốn tìm hiểu về một mức giá hợp lý tùy vào mục đích của mình.

Hình 2: Khi bấm vào mục này, khách hàng sẽ có thể nhập các thông số về hãng xe, dòng xe, số km đã đi, khu vực muốn bán.



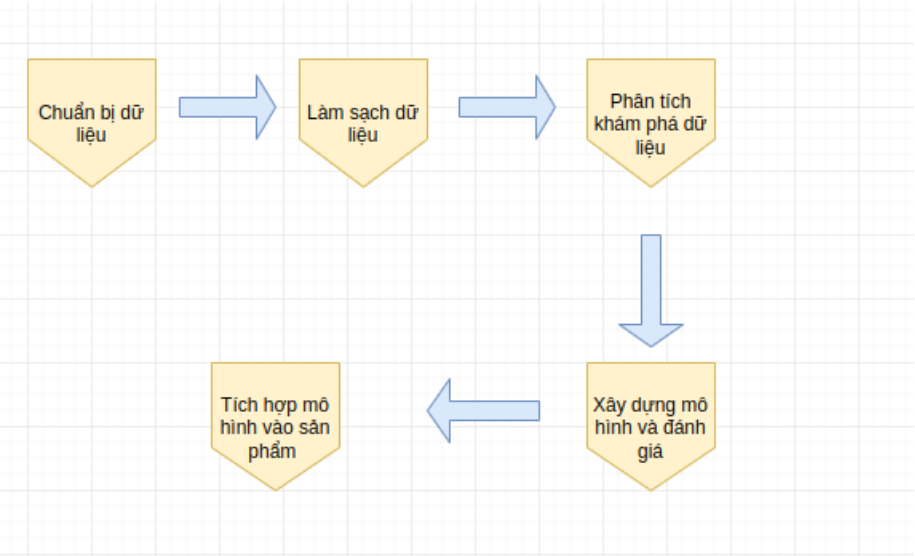
Hình 3: Sau đó, chỉ cần nhấn vào phần ĐỊNH GIÁ XE, sẽ hiện ra giá xe dự đoán của chiếc xe đó.

Nhóm 4 chúng tôi cũng sẽ xây dựng nên một hệ thống có khả năng dự đoán giống như hệ thống "Định giá xe cũ"của trang xe Chợ tốt để giải quyết vấn đề: Giúp cho khách hàng, những người mua - bán xe, có được một "nhà tư vấn tự động", đưa ra mức giá tư vấn cho họ để giúp họ có những quyết định hợp lý, phù hợp với lợi tích của chính bản thân mình.

2 Nội dung dự án

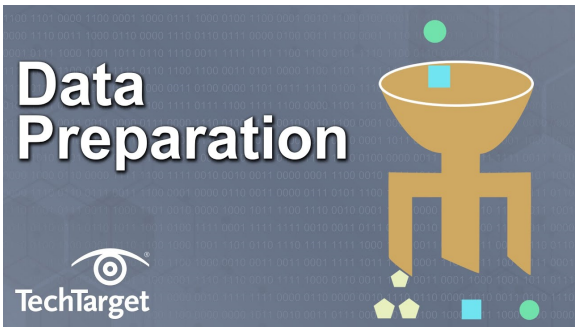
Vấn đề đã có, ý tưởng giải quyết vấn đề cũng đã có, như đã đề cập ở trên đó là xây dựng hệ thống học máy để dự đoán giá xe thông qua giao diện của một trang web. Bây giờ, chúng ta cùng xây dựng nội dung cho ý tưởng của dự án này.

Trước tiên, ta nhận thấy, đây là một dự án thực tế có áp dụng những công cụ học máy. Chính vì vậy ta chuẩn hóa nội dung của của dự án này về quy trình thực hiện một dự án Khoa học dữ liệu - Học máy hoàn chỉnh, gồm có 5 bước, được thể hiện qua hình vẽ dưới đây.



2.1 Bước 1: Chuẩn bị dữ liệu (Data Preparation)

Học máy được xây dựng nhờ nền tảng dữ liệu, chính vì vậy, bước đầu tiên, ta cần phải chuẩn bị một nền tảng dữ liệu. Nhóm chúng tôi thực hiện việc chuẩn bị dữ liệu thông qua việc thu thập các dữ liệu công khai từ trang web của Chợ tốt (https://xe.chotot.com/mua-ban-xe-may). Dữ liệu sẽ là dạng có cấu trúc, gồm 8 trường thuộc tính: Hãng xe, Năm sản xuất, tình trạng xe, số km đã đi, dòng xe, loại xe, dung tích xe, giá xe. Trong đó, đặc biệt quan trọng nhất sẽ là giá xe, đây là trường thuộc tính đầu ra của mô hình học máy, 7 trường thuộc tính kia sẽ là các đầu vào của mô hình.

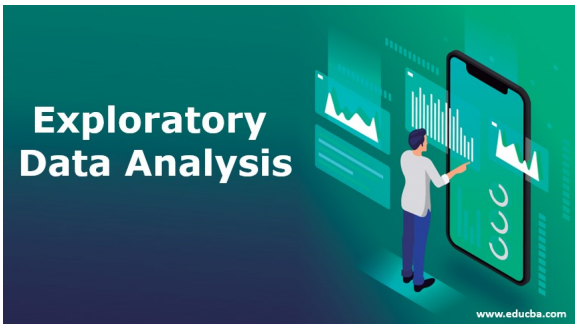


2.2 Bước 2: Làm sạch dữ liệu (Data Cleaning)

Đây là một bước vô cùng quan trọng, dữ liệu khi chúng ta thu thập có thể bị mất, bị thiếu, không có tính nhất quán. Việc làm sạch dữ liệu sẽ giúp chúng ta đảm bảo dữ liệu được đầy đủ, không bị thiết, đảm bảo được tính nhất quán. Có như vậy thì mô hình của chúng ta mới có thể hoạt động một cách trơn tru nhất.

2.3 Bước 3: Phân tích - Khám phá dữ liệu (Exploratory Data Analysis)

Có hai dạng phân tích khám phá dữ liệu, ứng với các chức năng nghề nghiệp Data Analyst (Chuyên viên phân tích dữ liệu) và Data Scientist (Chuyên viên Khoa học dữ liệu). Đối với vị trí Data Analyst, việc phân tích thường bắt đầu với những câu hỏi, sau đó sẽ đi sâu vào khía cạnh dữ liệu, tìm ra những thông tin hữu ích để trả lời cho câu hỏi đó. Cùng với đó, họ cũng sẽ đi sâu



vào những khía cạnh khác, những giá trị ẩn sâu bên trong "đống dữ liệu"đang chờ được khám phá, từ đó họ có thể trích rút ra những thông tin hữu ích cho tổ chức, doanh nghiệp của họ. Data Scientist thì thiên về hướng thứ hai, đi sâu vào trong "đống dữ liệu để tìm thông tin hữu ích, nhưng do đặc thù công việc không giống như Data Analyst, chuyên sâu về mảng phân tích, suy diễn. Công việc chính của Data Scientist là xây dựng mô hình, nên các Data Scientist thường tìm những thông tin hữu ích có thể giúp ích cho mô hình của mình, hiểu được những thuộc tính đầu vào nào sẽ có sự tác động nhất định đến thuộc tính đầu ra, từ đó có sự chuẩn bị về những kịch bản có thể có trong mô hình.

Về phần này, việc phân tích dữ liệu, khám phá về bộ dữ liệu xe máy sẽ được nhóm 4 đi theo hướng của các Data Scientist, tập trung đào sâu vào bộ dữ liệu sau khi đã làm sạch, tìm hiểu sự phụ thuộc của các thuộc tính với nhau, và sự ảnh hưởng của các thuộc tính đầu vào đối với giá trị đầu ra.

2.4 Bước 4: Xây dựng mô hình và đánh giá (Modeling / Building model and Evaluating)

Sau khi đã có cái nhìn tổng quát, nhóm 4 chúng tôi sẽ đi xây dựng, kết hợp

với sự đánh giá các mô hình này để lựa chọn ra mô hình tốt nhất. Các thuật toán xây dựng mô hình được lựa chọn là những công cụ mạnh mẽ, phù hợp với bài toán hồi quy (Dự đoán giá xe máy thuộc lớp bài toán hồi quy trong Học có giám sát của mô hình Học máy), bao gồm: Hồi quy tuyến tính, Rừng ngẫu nhiên, Mạng Nơron.

2.5 Bước 5: Tích hợp mô hình vào sản phẩm

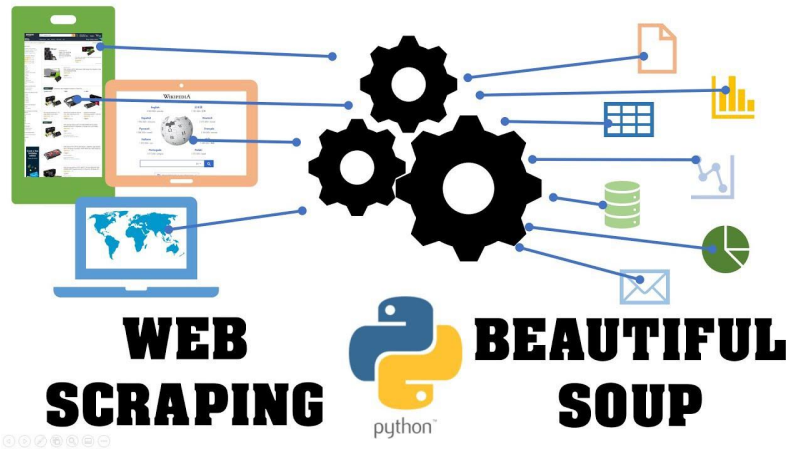
Cuối cùng, ta sẽ xây dựng một trang giao diện web, cho phép người dùng nhập các các trường dữ liệu như ở bước 1 và đưa ra dự đoán.

Trên đây là tóm tắt về 5 bước thực hiện ý tưởng của dự án, ta sẽ xem chúng như là 5 mô-đun, với mô-đun cuối cùng chính là sản phẩm. Sau đây ta sẽ bắt tay vào làm cụ thể, đưa ra chức năng chính của từng mô-đun.

3 Công cụ thực hiện

Công cụ chủ đạo của dự án này sẽ là ngôn ngữ lập trình Python, cụ thể:

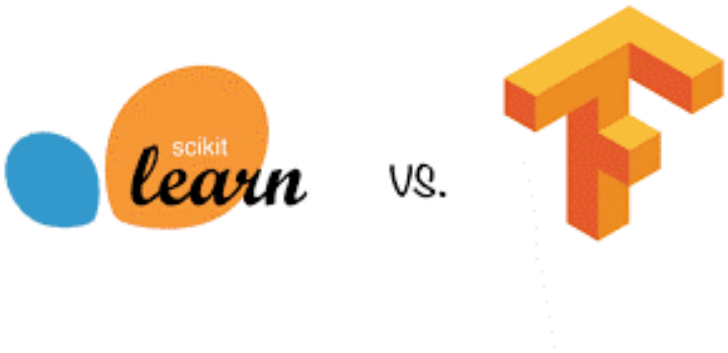
Đối với mô-đun thứ nhất, chuẩn bị dữ liệu, ta sẽ sử dụng thư viện BeautifulSoup cùng với thư viện Request để cào dữ liệu từ trang Chợ tốt về và lưu vào một tập tin đuôi .csv chứa dữ liệu thô



Đối với hai mô-đun tiếp theo, sẽ là bộ các thư viện rất phổ biến: Numpy, Pandas, Matplotlib (gồm cả Seaborn) để làm sạch và phân tích dữ liệu

Với mô-đun thứ 4, xây dựng mô hình, sẽ sử dụng tới thư viện Scikit-learn,

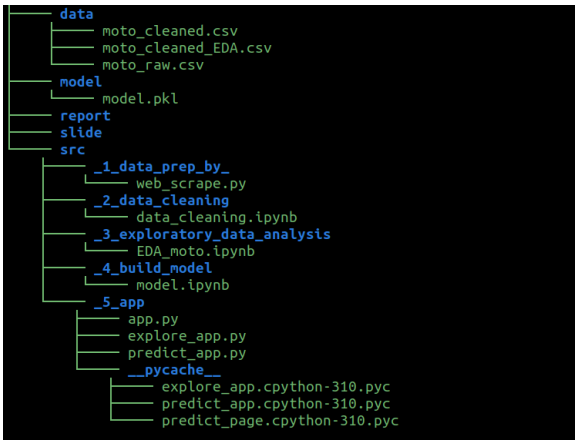
Xgboost và framework rất nổi tiếng cho học sâu, đó là Tensorflow



Cuối cùng, để xây dựng sản phẩm (mô-đun cuối), chúng tôi sẽ sử dụng một công cụ rất mạnh của Python, nó hỗ trợ việc tạo ra giao diện website một cách đẹp và vô cùng đơn giản, giúp cho những nhân lực trong ngành Khoa học dữ liệu có thể kiểm thử và giới thiệu mô hình của mình thông qua trang web một cách nhanh gọn mà không cần phải sử dụng đến các công cụ truyền thống như HTML, CSS và JavaScript, và thư viện đấy có tên gọi là Streamlit

4 Kết quả chính của các mô-đun và chức năng sản phẩm

Để xây dựng các mô-đun, trước hết cần xây dựng một thư mục chứa toàn bộ dự án, sau đó sẽ phân ra trong thư mục con là các mô-đun, cụ thể thư mục được tạo ra như sau:



Trong đó, thư mục /scr là thư mục chứa toàn bộ mã nguồn liên quan đến việc thực thi các mô-đun, đặt tên với các số và tên tương ứng, thư mục /model để lưu mô hình ở mô-đun 4, thư mục /data chứa các dữ liệu (thô, đã được làm sạch), thư mục /report sẽ dùng để chứa chính báo cáo này, và /slide để chiếu tập tin phục vụ cho việc trình chiếu.

4.1 Chuẩn bị dữ liệu

Việc chuẩn bị dữ liệu được tạo từ tập tin *web$_s$crape.py* ở mô-đun 1, nội dung chính trong tập tin này được tóm tắt như sau: Thư viện Request sẽ đưa các phương thức giao tiếp đến trang web cần cào dữ liệu (Cụ thể ở đây là Chợ tốt), sau đó thư viện Beautiful Soup sẽ lấy ra nội dung tương ứng ở các thẻ HTML của trang web này. Thực hiện vòng lặp qua các trang con của trang web chính, và tìm các thẻ HTML, các tên class, id mang thông tin các trường thuộc tính chúng ta cần, và "cào"chúng xuống, lưu lại vào tập dữ liệu thô trong thư mục data.

Đây là hình ảnh minh họa về dữ liệu thu được.

```
Suzuki,1996,Đãsửdụng,Null,Sport/Xipo,1234,Taycôn/Moto,25.500.000đ
Piaggio,2014,Đãsửdụng,100-175cc,Vespa,11000,Tayga,32.500.000đ
Yamaha,2013,Đãsửdụng,100-175cc,Exciter,25,Taycôn/Moto,27.500.000đ
Yamaha,2015,Đãsửdụng,100-175cc,Exciter,37,Xesô,26.500.000đ
Honda,1999,Đãsửdụng,100-175cc,Dream,25,Xesô,25.000.000đ
Honda,2019,Đãsửdụng,100-175cc,Cub,585869,Xesô,3.300.000đ
Yamaha,2013,Đãsửdụng,Null,Sirius,30,Xesô,7.800.000đ
Honda,2019,Đãsửdụng,Null,Winner,1000,Taycôn/Moto,16.500.000đ
Honda,2020,Đãsửdụng,100-175cc,Vario,1,Tayga,40.500.000đ
Yamaha,2018,Đãsửdụng,Null,Nvx,46000,Tayga,26.000.000đ
Honda,2009,Đãsửdụng,100-175cc,AirBlade,15000,Tayga,12.500.000đ
Honda,2021,Đãsửdụng,100-175cc,AirBlade,1000,Tayga,35.000.000đ
Honda,2019,Đãsửdụng,Null,Winner,1000,Taycôn/Moto,20.800.000đ
Honda,2016,Đãsửdụng,100-175cc,SHMode,12,Tayga,46.000.000đ
Honda,2012,Đãsửdụng,100-175cc,Vision,1,Tayga,13.800.000đ
Honda,2022,Đãsửdụng,100-175cc,WinnerX,3000,Taycôn/Moto,25.000.000đ
Honda,2011,Đãsửdụng,100-175cc,AirBlade,355244,Tayga,14.700.000đ
Honda,2013,Đãsửdụng,100-175cc,AirBlade,15000,Tayga,13.000.000đ
Honda,2017,Đãsửdụng,Null,Cub,20000,Xesô,8.800.000đ
Piaggio,2012,Đãsửdụng,100-175cc,LX,25689,Tayga,12.500.000đ
Yamaha,2016,Đãsửdụng,Null,Exciter,18000,Xesô,23.500.000đ
Honda,2018,Đãsửdụng,100-175cc,SH,30000,Tayga,86.000.000đ
Honda,2017,Đãsửdụng,100-175cc,Wave,18000,Xesô,14.800.000đ
Honda,2012,Đãsửdụng,100-175cc,AirBlade,5,Tayga,15.500.000đ
Honda,2019,Đãsửdụng,100-175cc,Winner,1,Taycôn/Moto,27.900.000đ
```

4.2 Làm sạch dữ liệu

Trong một quy trình xây dựng một hệ thống học máy, làm sạch dữ liệu là bước thứ 3 sau khi dự án đã được lên ý tưởng và dữ liệu thô đã được thu thập.

Việc làm sạch dữ liệu thô giúp cho dữ liệu được cập nhật một cách chính xác, rõ ràng, minh bạch, đảm bảo nguồn dữ liệu của chúng ta không bị hỏng, không bị thiếu trước khi được đưa vào quy trình phân tích, hay chuẩn hóa trước khi đưa vào xây dựng mô hình Học máy.

Vì vậy có thể khẳng định, làm sạch dữ liệu là một quy trình hết sức quan trọng. Nếu không được làm sạch, sẽ dẫn đến việc báo cáo của chúng ta bị sai lệch, tệ hơn nữa là mô hình học máy đưa ra kết quả không đáng tin cậy. Ảnh hưởng xấu tới quá trình ra quyết định. Do đó, cần phải thực hiện công việc làm sạch dữ liệu một cách triệt để, đảm bảo đủ các tính chất cần thiết của một "bộ dữ liệu sạch".

Trong mô-đun 2, ta tiến hành làm sạch bộ dữ liệu từ tập notebook của Python, toàn bộ quá trình làm sạch bộ dữ liệu được thể hiện trong tập của mô-đun 2. Kết quả sau khi dữ liệu được làm sạch như sau:

```
Hang_xe,Tuoi_xe,Nam_dang_ky,Tinh_trang_xe,Dung_tich_xe,Dong_xe,So_km_da_di,Loai_xe,Gia_xe
Suzuki,27,1996,Đã sử dụng,Không rõ,Sport/Xipo,1234,Tay côn,25500000.0
Piaggio,9,2014,Đã sử dụng,100 - 175 cc,Vespa,11000,Tay ga,32500000.0
Yamaha,10,2013,Đã sử dụng,100 - 175 cc,Exciter,25,Tay côn,27500000.0
Yamaha,8,2015,Đã sử dụng,100 - 175 cc,Exciter,37,Xe số,26500000.0
Honda,24,1999,Đã sử dụng,100 - 175 cc,Dream,25,Xe số,25000000.0
Honda,4,2019,Đã sử dụng,100 - 175 cc,Cub,585869,Xe số,3300000.0
Yamaha,10,2013,Đã sử dụng,Không rõ,Sirius,30,Xe số,7800000.0
Honda,4,2019,Đã sử dụng,Không rõ,Winner,1000,Tay côn,16500000.0
Honda,3,2020,Đã sử dụng,100 - 175 cc,Vario,1,Tay ga,40500000.0
Yamaha,5,2018,Đã sử dụng,Không rõ,Nvx,46000,Tay ga,26000000.0
Honda,14,2009,Đã sử dụng,100 - 175 cc,AirBlade,15000,Tay ga,12500000.0
Honda,2,2021,Đã sử dụng,100 - 175 cc,AirBlade,1000,Tay ga,35000000.0
Honda,4,2019,Đã sử dụng,Không rõ,Winner,1000,Tay côn,20800000.0
Honda,7,2016,Đã sử dụng,100 - 175 cc,SHMode,12,Tay ga,46000000.0
Honda,11,2012,Đã sử dụng,100 - 175 cc,Vision,1,Tay ga,13800000.0
Honda,1,2022,Đã sử dụng,100 - 175 cc,WinnerX,3000,Tay côn,25000000.0
Honda,12,2011,Đã sử dụng,100 - 175 cc,AirBlade,355244,Tay ga,14700000.0
Honda,10,2013,Đã sử dụng,100 - 175 cc,AirBlade,15000,Tay ga,13000000.0
Honda,6,2017,Đã sử dụng,Không rõ,Cub,20000,Xe số,8800000.0
Piaggio,11,2012,Đã sử dụng,100 - 175 cc,LX,25689,Tay ga,12500000.0
Yamaha,7,2016,Đã sử dụng,Không rõ,Exciter,18000,Xe số,23500000.0
Honda,5,2018,Đã sử dụng,100 - 175 cc,SH,30000,Tay ga,86000000.0
Honda,6,2017,Đã sử dụng,100 - 175 cc,Wave,18000,Xe số,14800000.0
Honda,11,2012,Đã sử dụng,100 - 175 cc,AirBlade,5,Tay ga,15500000.0
```

4.3 Phân tích, khám phá dữ liệu

Một trong những bước cần thiết trước khi chúng ta bước vào quá trình xây dựng mô hình học máy dự đoán giá xe máy đó là phân tích - khám phá dữ liệu của chúng ta, đưa ra được những thông tin hữu ích cơ bản từ tập giá trị này. Đồng thời, có thể kết hợp thêm việc làm sạch dữ liệu nếu thấy cần thiết.

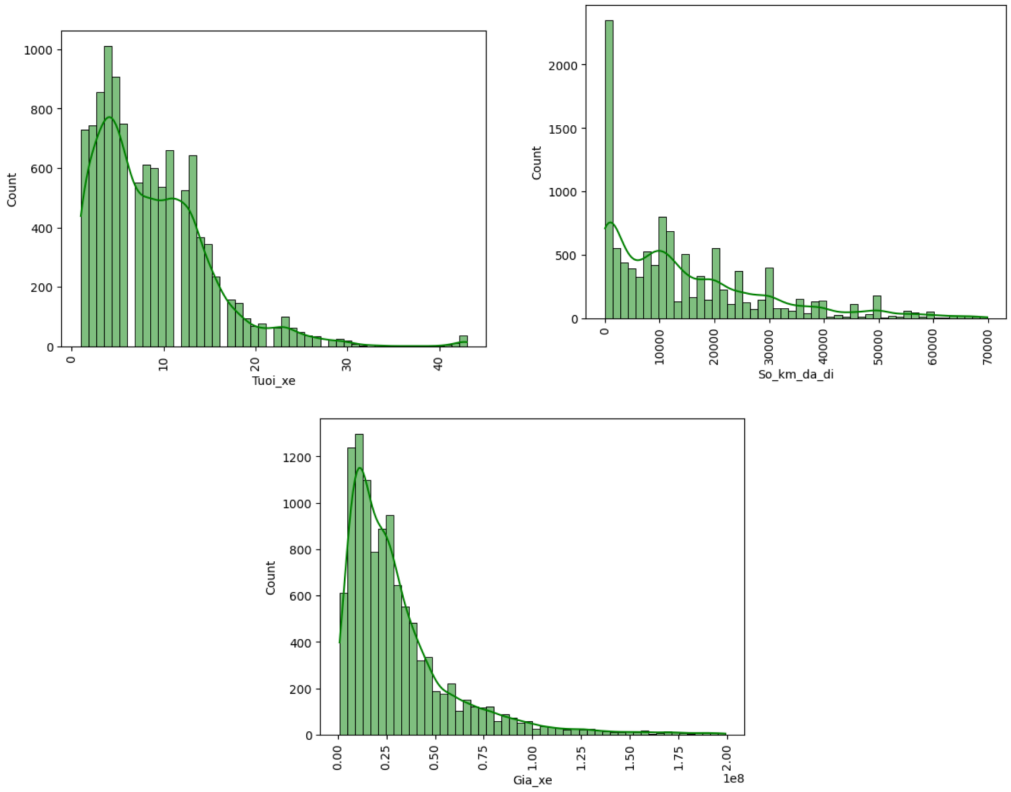
Việc có cái nhìn sơ bộ, đánh giá tổng quan ban đầu về mô hình là hết sức cần thiết và quan trọng, từ đó, giúp chúng ta định hướng ban đầu về các nhân tố tác động đến giá cả từ tập thuộc tính của dữ liệu, giúp đưa ra quyết định xây dựng mô hình một cách tốt hơn.

Từ việc phân tích - khám phá thực hiện trong mô-đun 3 được lưu vào một tập tin notebook, ta rút ra được một số kết luận phục vụ cho mô hình học máy:

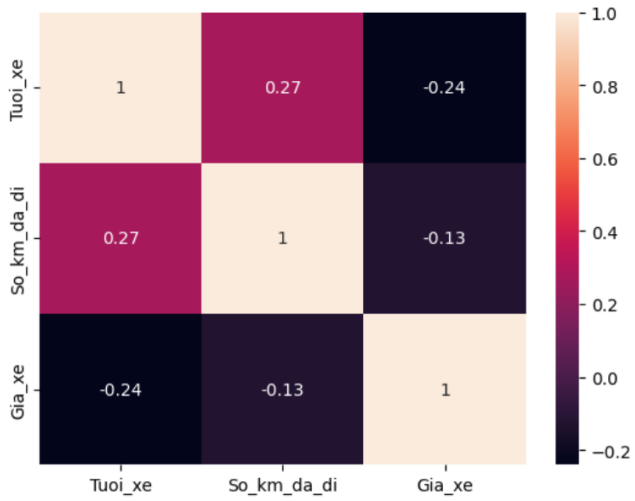
1. Có thể thấy có độ lệch "kha khá"về tham số giá trị trung bình trong trường thuộc tính so_km_da_di so với điểm trung vị, có thể điều này do sự xuất hiện của các điểm đột biến của thuộc tính so_km_da_di trong bộ dữ liệu, điều tương tự cũng xuất hiện với "Gia_xe"(Cần phải lưu ý các điểm này).

	Tuoi_xe	So_km_da_di	Gia_xe
count	12216.000000	12216.000000	1.221600e+04
mean	9.031189	30542.483219	3.787350e+07
std	6.548294	85909.979118	1.076080e+08
min	1.000000	0.000000	1.000000e+06
25%	4.000000	3000.000000	1.150000e+07
50%	8.000000	12000.000000	2.280000e+07
75%	13.000000	25000.000000	3.950000e+07
max	43.000000	1000000.000000	8.888888e+09

2. Cả ba trường dữ liệu định lượng đều phân bố chủ yếu ở những mức giá trị bên trái góc nhìn đối xứng của trục tung, như giá xe phân bố chủ yếu từ mức 75 triệu đồng trở xuống, số km đã đi phân bố nhiều ở mức dưới 30000km, còn tuổi xe đa phần ở mức từ 15, 16 tuổi trở xuống (Tức hầu hết các xe được sản xuất từ giai đoạn khoảng năm 2008 trở đi)

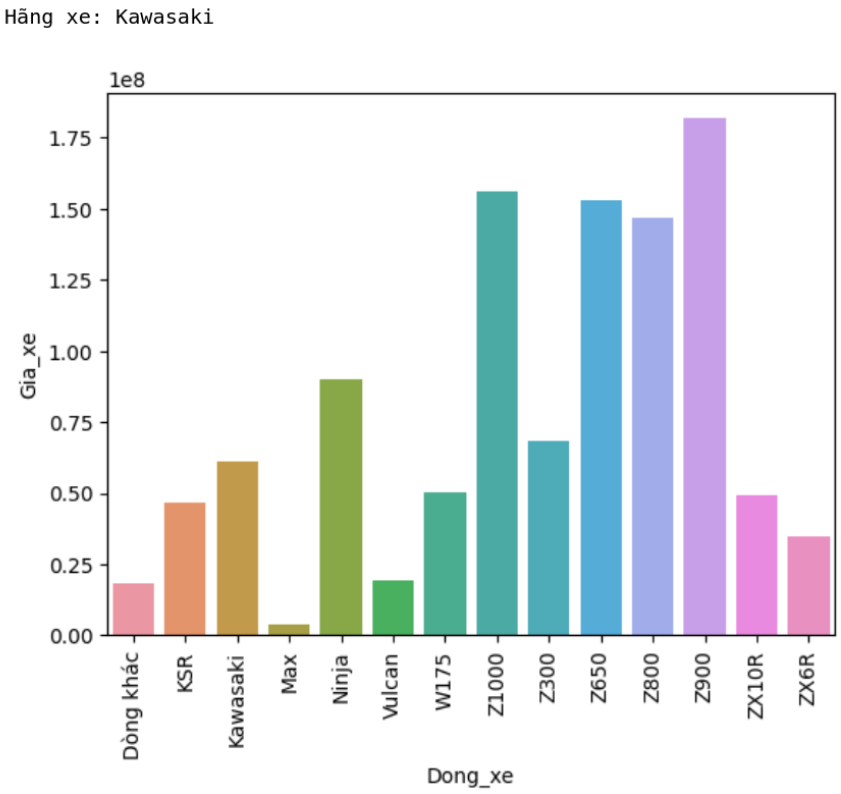


3. Hệ số tương quan giữa "Gia_xe"với hai trường thuộc tính trên có giá trị tuyệt đối không cao, thể hiện rằng rất rất ít mối quan hệ tuyến tính giữa các đại lượng này.



4. Biểu đồ cột cho ta thấy sự vượt trội về số lượng của các xe đước bán đến từ Honda, có lẽ đây cũng sẽ là hãng xe được giao bán nhiều nhất trên thị trường.

5. Hầu hết ở các hãng, các dòng đều có sự chênh lệch khá nhiều về giá trung bình, đồng thời, dòng xe luôn thuộc về một hãng xe, do đó khi xây dựng mô hình dự báo về giá xe, ta có thể loại đi thuộc tính về hãng xe. Một ví dụ minh họa:

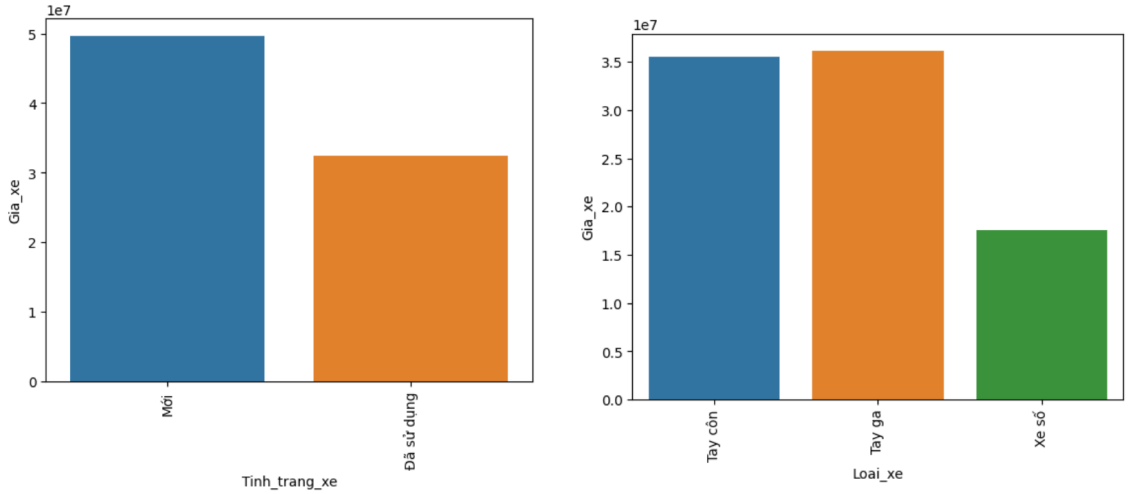


6. Về giá trung bình:

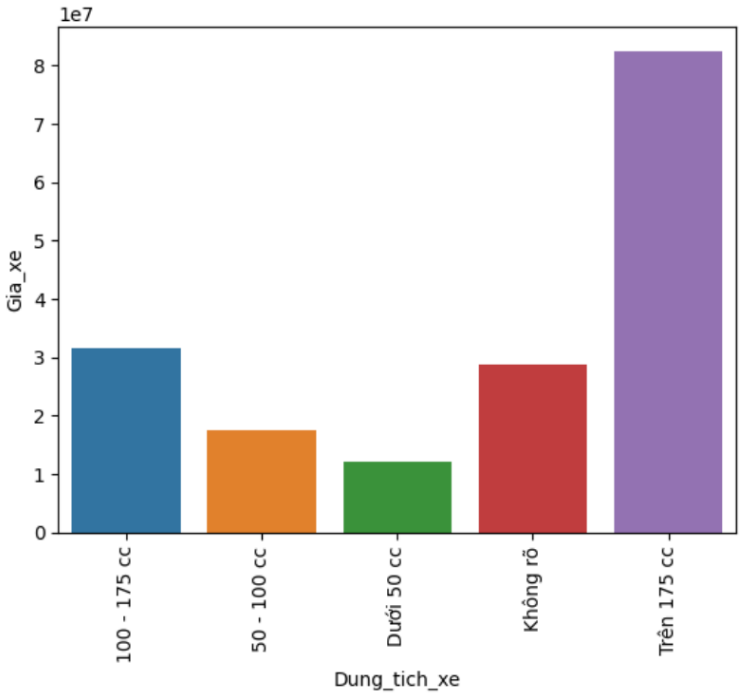
- Xe tay ga và xe tay côn có giá trung bình cao hơn so với xe số.
- Xe mới có giá cao hơn so với xe cũ cùng loại.
- Xe phân khối càng cao thì giá càng cao (Tỷ lệ thuận).

4.4 Xây dựng và đánh giá mô hình

Sau khi đã trải qua các bước, từ việc chuẩn bị dữ liệu, làm sạch, đưa ra những sự phân tích, đánh giá cơ bản, thì bây giờ chúng ta sẽ đến với bước



Giá xe trung bình theo thuộc tính: Dung_tich_xe



quan trọng nhất, cũng chính là bước chính thức của dự án này, đó là chúng ta sẽ xây dựng một mô hình dự đoán giá xe máy, có sự đánh giá, lựa chọn ra mô hình tốt nhất trước khi chúng ta đưa vào hệ thống sản phẩm (có giao diện website).

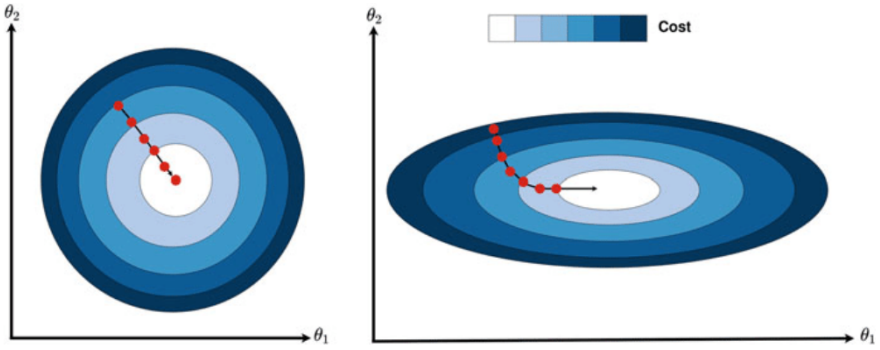
Bài toán lần này là một bài toán về hồi quy, dựa đoán giá trị thực của một thuộc tính, do đó chúng ta cần chọn ra những công cụ, những thuật toán phù hợp để xây dựng tốt mô hình.

Trong dự án này, sẽ sử dụng các công cụ hồi quy đang được hướng dẫn căn bản từ khóa học Machine Learning Specialization của công ty DeepLearning.AI, bao gồm Hồi quy tuyến tính (Linear Regression) (Không chỉ gồm hồi quy tuyến tính mà còn có thể sử dụng tới hồi quy đa thức (Polynomial Regression) - một biến thể của hồi quy tuyến tính, Kỹ thuật hiệu chỉnh thông qua Ridge Regression), Rừng ngẫu nhiên (Random Forest) và Mạng Neural với hàm kích hoạt ở lớp (layer) cuối cùng là hàm tuyến tính (linear-activation).

Về thuật toán chi tiết của các mô hình, các bạn có thể xem trong thư mục về tài liệu tham khảo.

Trước tiên, để xây dựng được mô hình dữ liệu, chúng ta cần chuẩn hóa dữ liệu. ta sẽ chuẩn hóa các dữ liệu của các thuộc tính đầu vào về khoảng [-1, 1] bằng cách chuẩn hóa theo giá trị trung bình (Mean Normalization) để đảm bảo cho các thuật toán tối ưu được hiệu quả

Ví dụ minh họa về tác dụng của việc thực hiện chuẩn hóa dữ liệu khi thực hiện thuật toán Gradient descent:



Sau khi thực hiện quá trình chuyển hóa, dữ liệu ban đầu chuyển thành:

	Tuoi_xe	Tinh_trang_xe	Dung_tich_xe	Dong_xe	So_km_da_di	Loai_xe	Gia_xe
0	0.434884	0.008274	0.487634	0.245907	-0.189822	-0.536874	25.5
1	0.006313	0.008274	-0.262366	0.319353	-0.050020	-0.036874	32.5
2	0.030122	0.008274	-0.262366	-0.228670	-0.207129	-0.536874	27.5
3	-0.017497	0.008274	-0.262366	-0.228670	-0.206957	0.463126	26.5
4	0.363456	0.008274	-0.262366	-0.296466	-0.207129	0.463126	25.0
5	0.030122	0.008274	0.487634	0.223308	-0.207057	0.463126	7.8
6	-0.112735	0.008274	0.487634	0.364551	-0.193171	-0.536874	16.5
7	-0.136544	0.008274	-0.262366	0.308054	-0.207472	-0.036874	40.5
8	-0.088925	0.008274	0.487634	0.076415	0.451011	-0.036874	26.0
9	0.125360	0.008274	-0.262366	-0.505506	0.007241	-0.036874	12.5

Tiếp theo ta sẽ tách bộ dữ liệu ra thành các tập huấn luyện và kiểm thử mô hình, trong mỗi tập đó sẽ bao gồm các thành phần input X - chứa các thuộc tính ngoại trừ giá cả, output y chứa thuộc tính về giá cả. Trong tập kiểm thử chúng ta sẽ tách ra thêm thành hai tập "tập validation"và tập kiểm thử "thực sự", tập này có ý nghĩa để chúng ta kiểm thử mô hình và cố gắng tối ưu hóa các tham số, điều này sẽ giúp cho tập kiểm thử "thực sự"của chúng ta là đại diện tốt nhất, mang tính ngẫu nhiên nhất để đánh giá mô hình.

```
# Kiểm tra qua số chiều của các tập train test

X_train.shape, X_test.shape, X_val.shape, y_train.shape, y_test.shape, y_val.shape
```

```
((8895, 6), (1112, 6), (1112, 6), (8895,), (1112,), (1112,))
```

Bước sau đó sẽ là bước quan trọng nhất, có thể nói như là trái tim của cả dự án: Xây dựng mô hình và đánh giá, lựa chọn ra mô hình tốt nhất có thể. Như đã nói ở phần giới thiệu, trong phần xây dựng mô hình lần này, chúng ta sẽ sử dụng ba phương pháp:

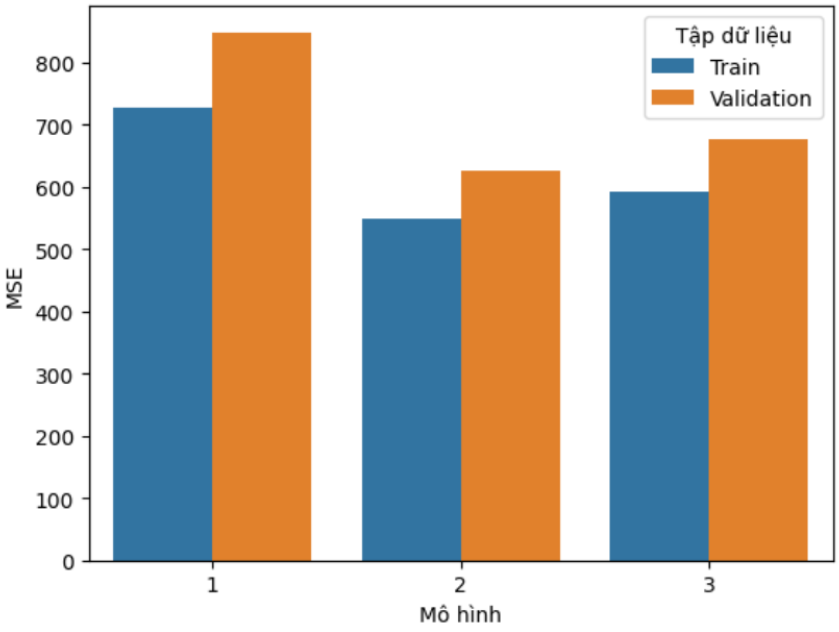
- Hồi quy tuyến
- Rừng ngẫu nhiên
- Mạng Neural

Để đánh giá tất cả các mô hình, ta sẽ sử dụng trung một tham số đó là MSE (Mean Squared Error), tham số đo trung bình về độ lệch bình phương giữa các giá trị dự đoán với giá trị thực tế của chúng.

Với phương pháp hồi quy tuyến tính, chúng ta sẽ đưa ra các mô hình sau:

- Mô hình đầu tiên của hồi quy tuyến tính:
  + Chúng ta sẽ áp dụng phương pháp hồi quy tuyến tính truyền thống, sử dụng thuật toán Gradient Descent để tìm ra tham số tối ưu cho việc dự đoán giá xe.
- Mô hình thứ hai của hồi quy tuyến tính:
  + Chúng ta sẽ thay hồi quy tuyến tính bằng hồi quy đa thức, với một đa thức bậc cao bất kỳ, việc lựa chọn này sẽ chọn đa thức bậc cao, để phục vụ cho mô hình thứ 4.
- Mô hình thứ ba của hồi quy tuyến tính:
  + Từ mô hình thứ hai, chúng ta sẽ thêm kỹ thuật hiệu chỉnh, hay còn được gọi là hồi quy Ridge để tránh bị Overfitting dữ liệu.

  Kết quả:

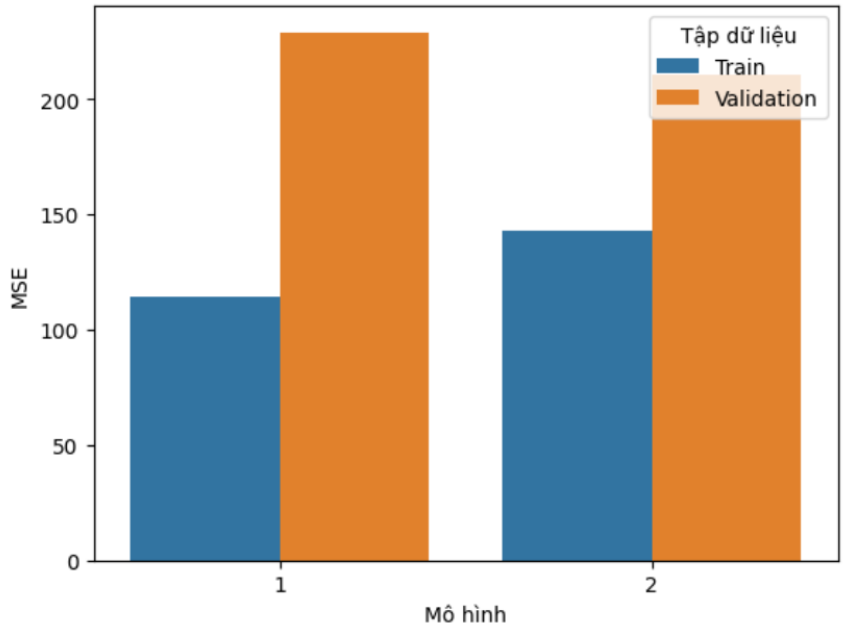


Với kết quả này, đối với phương pháp hồi quy tuyến tính truyền thống, ta sẽ chọn mô hình thứ hai.

Với phương pháp rừng ngẫu nhiên, chúng ta sẽ sử dụng hai phương pháp:

- Phương pháp rừng ngẫu nhiên truyền thống (Mô hình 1)
- XGboost - một biến thể của rừng ngẫu nhiên (Mô hình 2)

Cách chúng ta đánh giá cũng hoàn toàn tương tự như Hồi quy tuyến tính phía trên.



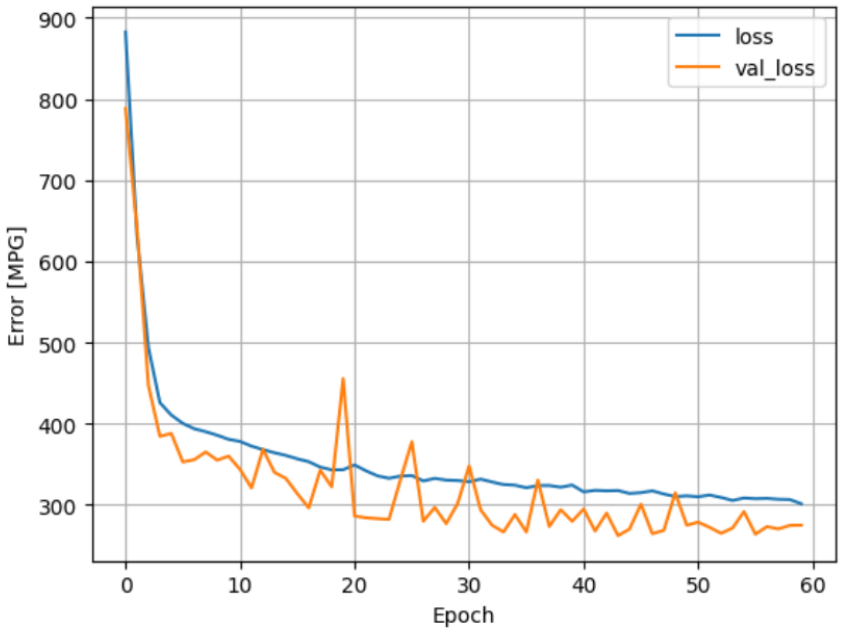
Kết quả đưa ta sự lựa chọn với XGboost (mô hình thứ 2)

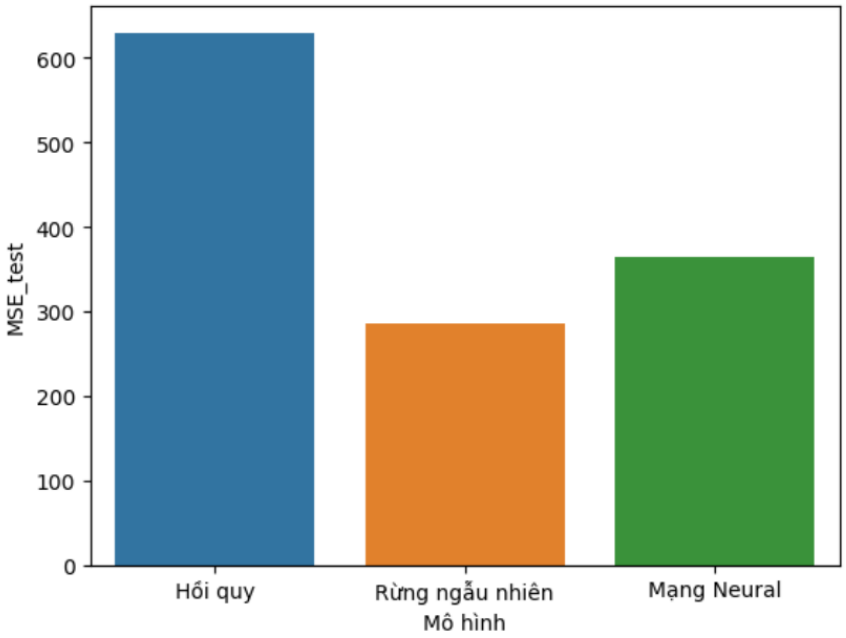
Cuối cùng, ta sẽ sử dụng một mạng Neural được xây dựng từ nền tảng Keras của TensorFlow, để tạo ra một mô hình so sánh với hai mô hình tốt nhất của hai phương pháp trên.

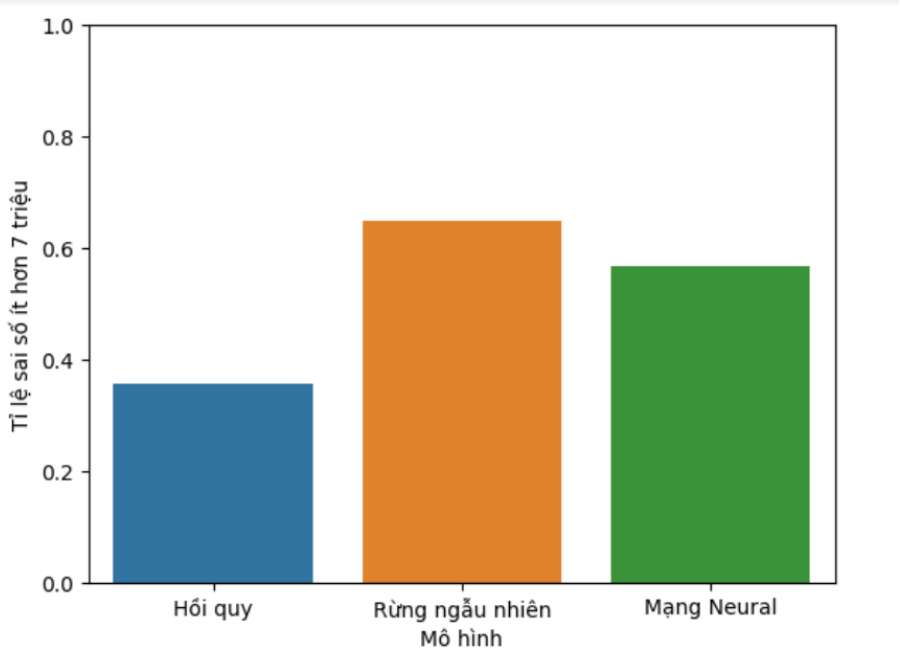
Đây là dữ liệu về MSE của mạng Neural đối với hai tập huấn luyện và validation.

Sau khi đã có 3 "đối thủ nặng ký"nhất, cuối cùng ta sẽ so sánh chúng với tập dữ liệu test thực sự, tập mà chúng ta đã tách ra hoàn toàn riêng biệt với tập validation, là tập dữ liệu "chuẩn"đại diện tốt cho dữ liệu test trong thực tế, do chúng không bị ta tác động như tập validation.

Từ các kết quả trên, so sánh về tỉ lệ giá xe dao động thấp hơn 7 triệu và MSE test, ta quyết định lựa chọn mô hình XGboost là mô hình cuối cùng để


loss
val_loss
Error [MPG]
Epoch
900
800
700
600
500
400
300
0
10
20
30
40
50
60

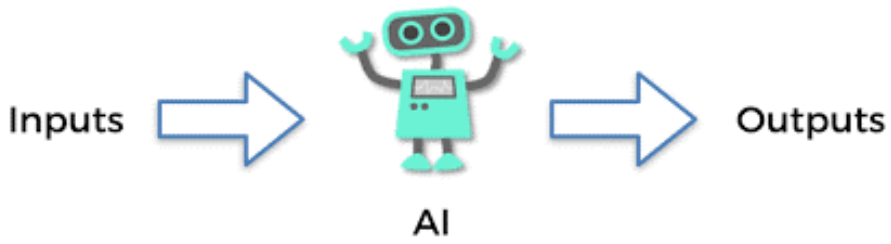

MSE_test
Mô hình
Hồi quy
Rừng ngẫu nhiên
Mạng Neural
600
500
400
300
200
100
0



tích hợp sản phẩm.

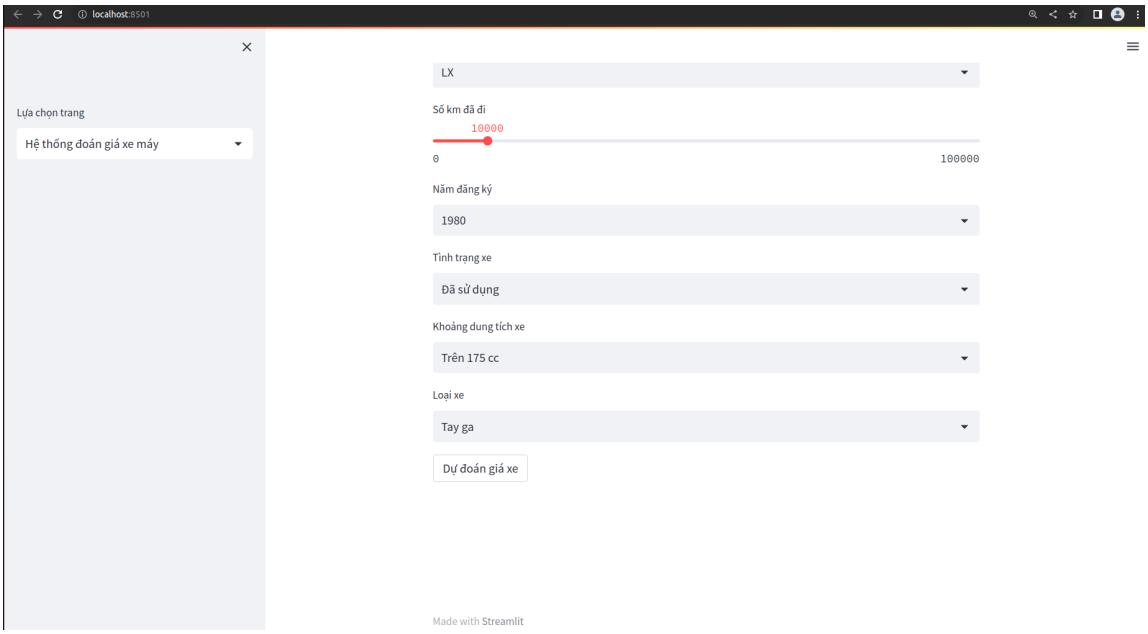
4.5 Chức năng chính của sản phẩm (Mô-đun cuối cùng)

Sau khi đã tạo ra được mô hình, ta lưu nó vào tập nhị phân (Sử dụng thư viện Pickle) trong thư mục /model, sản phẩm của chúng ta sẽ gọi đến file này, nó sẽ lưu thông tin của người dùng nhập vào, mang ý nghĩa đầu vào dành cho mô hình để dự đoán ra giá của chiếc xe máy. Như vậy có thể nói, sản phẩm của chúng ta có bản chất là một hệ thống Học máy - Trí tuệ nhân tạo để dự đoán ra giá của chiếc xe

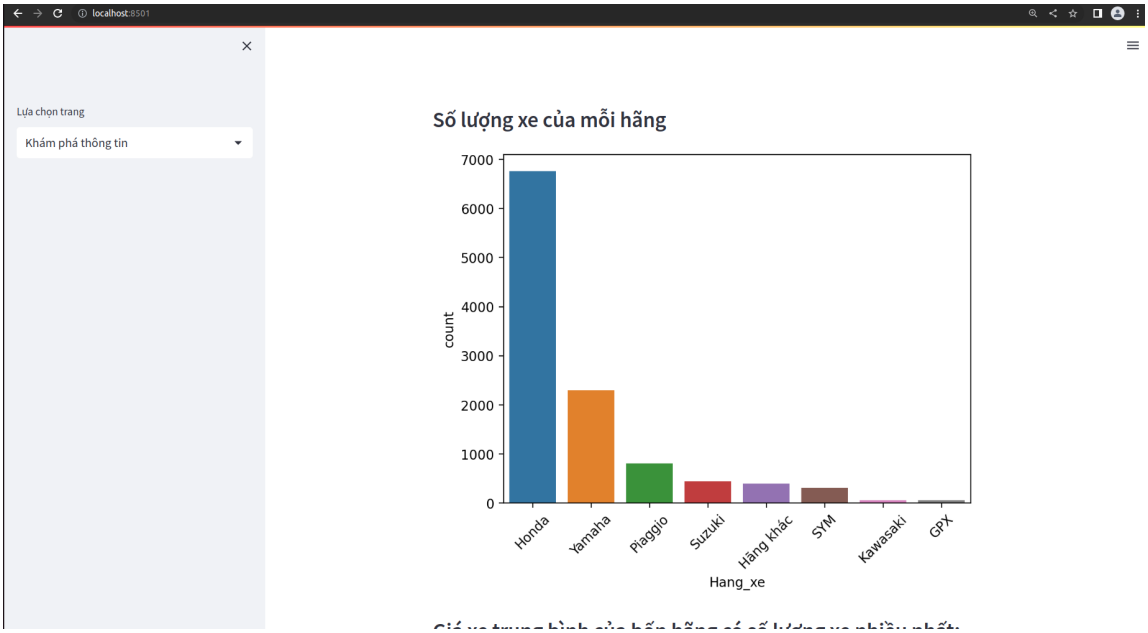


Sản phẩm có bản chất là một trang web, trong trang web này có hai trang con:

- Một trang đưa ra hệ thống dự đoán giá xe máy, cho phép người dùng nhập vào các thông số thuộc tính, hệ thống học máy từ mô hình sẽ đưa ra dự đoán về giá một cách hợp lý để người dùng tham khảo.
- Một trang khám phá, cho phép người dùng xem biểu đồ thông số của các hãng xe, và giá trung bình của các dòng xe trong số các hãng xe được giao bán nhiều nhất. Đây cũng là thông tin quan trọng, nhóm chúng tôi không chỉ hướng đến một dự án đưa ra dự đoán đơn thuần, mà còn muốn mang đến cho người dùng cái nhìn khách quan để đưa ra quyết định hợp lý nhất dựa vào biểu đồ.



Hình 4: Hình ảnh về trang đưa ra dự đoán giá xe



Hình 5: Hình ảnh ở trang khám phá

5 Tổng kết

Sau một quá trình thực hiện dự án, nhóm chúng tôi đã hoàn thành dự án. Mặc dù đã hoàn thành, nhưng nhận thấy dự án vẫn còn những điểm hạn chế cần phải cải thiện như:

- Giá trị MSE vẫn không thực sự quá nhỏ ở mô hình XGboost (Mô hình tốt nhất đã được chọn).
- Chưa thể phân tích hết được nhiều mặt khác trong quá trình Phân tích - Khám phá dữ liệu (Mô-đun 3).
- Chưa lấy được trường thuộc tính về thành phố (Nơi xe được bán) - một thuộc tính có thể cũng rất hữu ích.

Tuy vậy, hệ thống hoạt động cũng khá tốt, với nhiều trường hợp đối với các dòng xe phổ biến hiện nay như Dream, Wave, Sirius, Vespa, Vision, SH, Air-Blade,... hệ thống đưa ra được một mức giá cũng rất hợp lý theo mặt bằng chung, giúp cho người dùng có thể dựa vào đó để tham khảo về mức giá tùy cho mục đích của mình.

Những điểm phát triển tiếp theo:

- Thêm mới nguồn dữ liệu, có thể bổ sung thêm nguồn dữ liệu đa dạng hơn về các loại xe (Nguồn dữ liệu ban đầu thể hiện sự chiếm ưu thế của Honda).
- Phân tích sâu hơn về các khía cạnh mới (Ví dụ như sự phụ thuộc về giá giữa các tổ hợp thuộc tính thay vì chỉ một loại thuộc tính).
- Phát triển thêm về mặt sản phẩm, có thể sau khi đưa ra dự đoán, sẽ cho ra một số mẫu xe và mức giá có độ tương đồng với xe mà người dùng nhập vào, để họ có thêm sự tham khảo (Điều này có thể được hỗ trợ bởi thuật toán KNN).

Tài liệu

[1] Thư viện pandas, https://pandas.pydata.org/

[2] Thư viện numpy, https://numpy.org/

[3] Thư viện matplotlib, https://matplotlib.org/

[4] Thư viện scikit-learn, https://scikit-learn.org/stable/

[5] Framework tensorflow, https://www.tensorflow.org/

[6] Thư viện streamlit, https://streamlit.io/

[7] Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, *Aurélien Géron*

[8] Deep Learning with Python, *François Chollet*

[9] Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python, *Andrew Bruce, Peter C. Bruce, Peter Gedeck*