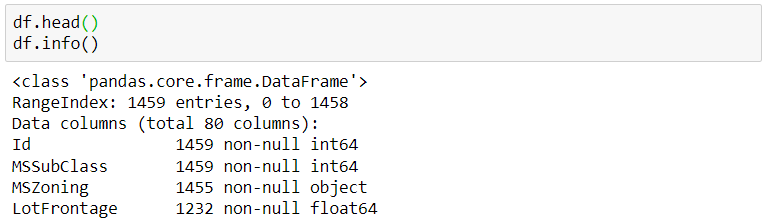
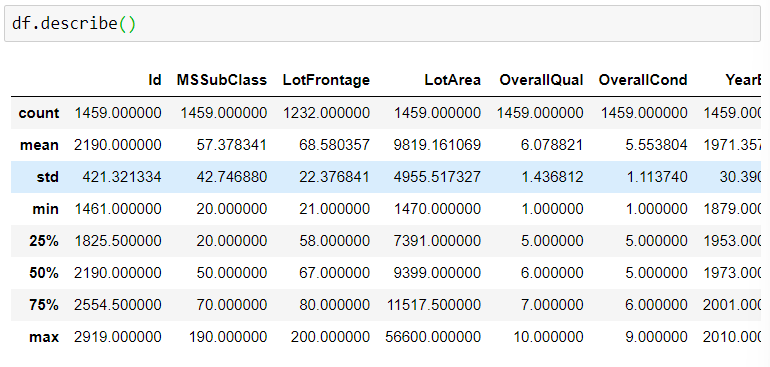
IV. Thực nghiệm

4.1. Mô tả dữ liệu

4.1.1. Tổng quan dữ liệu



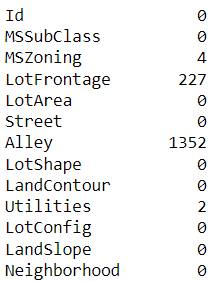
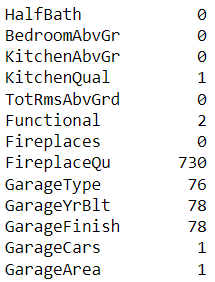
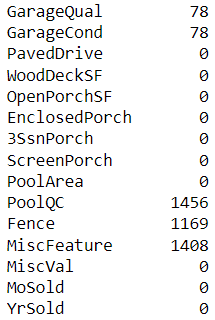
Dữ liệu chứa 80 thuộc tính và 1459 bản ghi



Miêu tả dữ liệu cho thấy dữ liệu tất cả đều mang giá trị dương (>0) nên có thể chính xác về mặt giá trị. Nhưng vẫn còn nhiều trường dữ liệu có giữ liệu trống hoặc không liên quan cần loại bỏ.

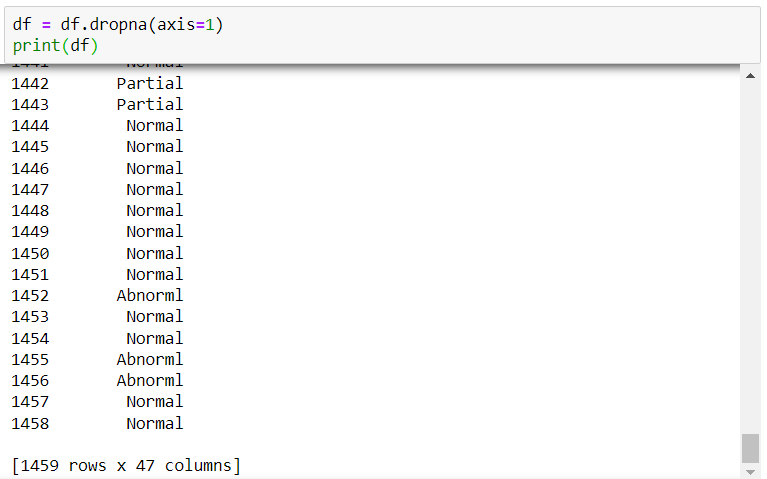
4.1.2. Các vấn đề của dữ liệu



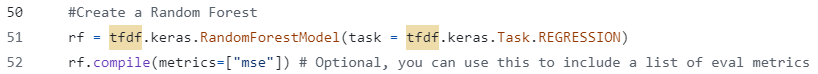
  

Vì là tệp dữ liệu về nhà ở nên ở mỗi trường dữ liệu đều có thêm các dữ liệu phân loại cụ thể, trong đó có rất nhiều dữ liệu trống và không cần thiết

4.2. Tiền xử lí dữ liệu



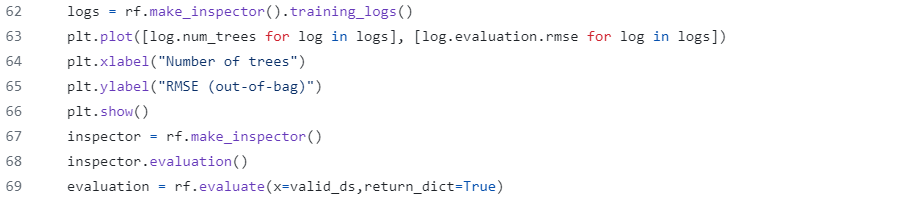
Trước tiên, trong tệp dữ liệu có chứa các trường dữ liệu tồn tại giá trị null, chúng ta cần xác định dữ liệu đó có thật sự cần thiết và loại bỏ khỏi tệp dữ liệu. Sau khi xóa các cột chứa giá trị null, tệp dữ liệu còn lại 47 thuộc tính cùng 1459 bản ghi.

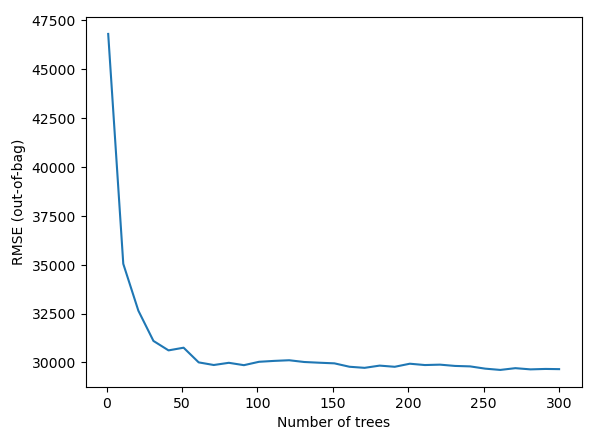


Trong trường hợp này, sử dụng công việc REGRESSION cho biết mô hình bài toán sẽ được dùng để dự đoán giá trị liên tục (hồi quy). Sau khi tạo mô hình thành công, trong quá trình “compile” sẽ chỉ định tham số huấn luyện và các độ đo để đánh giá hiệu suất của mô hình. Độ đo phổ biến được sử dụng trong các bài toán hồi quy để đo lường sự khác biệt giữa giá trị dự đoán và giá trị thực tế là “mse” (Mean Squared Error)



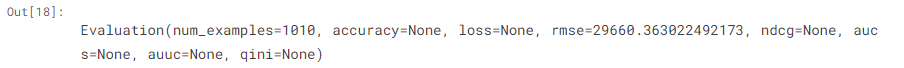
Sau khi đã tạo mô hình, mô hình sẽ được huấn luyện bằng dữ liệu đầu vào “train\_ds”. Dữ liệu huấn luyện này sẽ được sử dụng để điều chỉnh các tham số của mô hình Random Forest sao cho mô hình có thể dự đoán đúng các đầu ra tương ứng với các đầu vào.





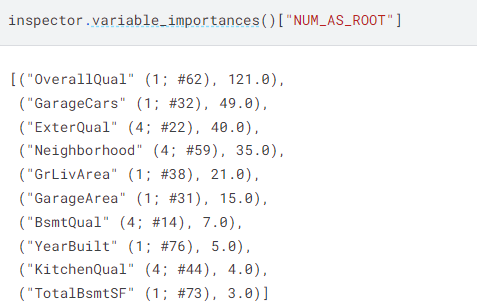
Biểu đồ ở trên cho thấy sự thay đổi các giá trị RMSE (Root Mean Square Error) khi them hoặc loại bỏ số lượng cây trong mô hình. Việc thêm nhiều cây có thể giúp tăng thêm hiệu suất nhưng cũng có thể gây ra overfitting trên tập dữ liệu huấn luyện

RMSE là một phép đo đánh giá mức độ chênh lệch giữa các giá trị dự đoán và các giá trị thực tế. Trong trường hợp này, RMSE được đánh giá từ nhật ký huấn luyện của mô hình Random Forest. Điểm số RMSE càng thấp, tức là mô hình dự đoán càng chính xác. Nếu có một điểm cụ thể của số lượng cây mà giá trị RMSE đạt đến mức thấp nhất hoặc ổn định, chúng ta có thể xác định được số lượng cây tốt nhất cho mô hình của mình



Số lượng ví dụ trong tập dữ liệu thử nghiệm được sử dụng để đánh giá mô hình là 1010. Với chỉ số accuracy của mô hình là None, có nghĩa là mô hình không được đánh giá dựa trên độ chính xác. Tương tự như giá trị mất mát (loss) cũng là None.

Giá trị của Root Mean Square Error (RMSE) của mô hình trên tập dữ liệu thử nghiệm là rmse=29660.363022492173. Ngoài ra các chỉ số phụ đều đưa kết quả None có nghĩa là mô hình này không được đánh giá dựa trên các chỉ số đó.



Đây là kết quả của biến có thứ tự quan trọng ví dụ như OverallQual là vị trí 1 trong danh sách các biến. Số 62 trong ngoặc đơn là chỉ số của biến này trong tập dữ liệu hoặc mô hình. Mức độ quan trọng của biến này được đo bằng giá trị 121.0.

OverallQual (Chất lượng tổng thể của căn nhà): Đây thường là một trong những biến quan trọng nhất khi dự đoán giá nhà. Mức độ quan trọng cao đề xuất rằng mô hình xem xét Chất lượng Tổng thể là yếu tố quan trọng trong việc dự đoán giá nhà. Mô hình có thể sử dụng biến này làm điểm bắt đầu để phân chia dữ liệu thành các nhóm nhỏ hơn dựa trên chất lượng tổng thể của căn nhà.

GarageCars (Số lượng xe trong ga-rơ): Đây cũng là một yếu tố quan trọng, vì số lượng xe có thể ảnh hưởng đến giá nhà. Một ga-rơ có thể được sử dụng để lưu trữ các phương tiện, và số lượng xe mà nó có thể chứa có thể ảnh hưởng đến giá trị của căn nhà.

ExterQual (Chất lượng bên ngoài): Chất lượng bên ngoài của căn nhà cũng có ảnh hưởng đáng kể đến giá trị của nó. Mô hình có thể sử dụng thông tin này để phân loại các căn nhà thành các nhóm có chất lượng bên ngoài khác nhau, từ đó giúp dự đoán giá trị của chúng.

Neighborhood (Khu vực): Khu vực mà căn nhà nằm trong có thể ảnh hưởng lớn đến giá nhà. Mô hình có thể sử dụng thông tin này để xác định các khu vực có giá trị cao hơn và thấp hơn, từ đó dự đoán giá trị của các căn nhà trong khu vực đó.

Và những biến khác như GrLivArea, GarageArea, BsmtQual, YearBuilt, KitchenQual, và TotalBsmtSF cũng có ảnh hưởng đến giá nhà và được mô hình coi là quan trọng trong quá trình dự đoán.