**Phân tích Data**

Trong bài toán đối với dự đoán giá nhà, trước hết chúng ta sẽ lập một biểu đồ để xem xét phân phối các giá nhà. Dựa ở BIỂU ĐỒ PHÍA DƯỚI, ta thấy phân phối trông giống với một phân phối chuẩn(Guassian Distribution). Phần lớn các ngôi nhà đều có giá trong khoảng từ $10000-$25000, và sau đó là một phần ít các nhà giá cao. Điều này được thể hiện rõ qua việc biểu đồ xuất hiện một cái ‘’đuôi’’ về phía tay phải; còn những ngôi nhà giá rẻ thì lại tập trung về phía tay trái.

1. Phân tích rời rạc với một số đặc trưng.

Tiếp theo, để tìm hiểu vì sao phân phối giá nhà lại có hình dạng như vậy, ta sẽ tìm hiểu mối liên hệ của giá nhà với một số đặc trưng khác(feature).

a) Đặc trưng được biểu hiện bằng số

Giả sử biểu đồ(scatterplot) giữa giá nhà và diện tích sống sau đây:

(BIỂU ĐỒ VỚI GrLivArea)

Dựa trên biểu đồ, giá nhà và diện tích có một mối tương quan tuyến tính với nhau. Điều này hoàn toàn hợp lí bởi vì diện tích nhà càng rộng thì khả năng cao là giá nhà cũng sẽ cao.

b) Đặc trưng đặc trưng phân loại.

Ví dụ như mối quan hệ giữa giá nhà và chất lượng của nhà(được đánh giá theo một thang từ 1-10). Hãy xem biểu đồ hộp(box-plot) ở phía dưới:

(BIỂU ĐỒ BOX)

Dựa trên biểu đồ ta thấy giá nhà và chất lượng có mối tương quan đúng như dự đoán: chất lượng càng cao, nhà càng đắt. Không những vậy, ta có thể thấy giá nhà lớn nhất, nhỏ nhất, trung vị của từng chất lượng nhà cũng có mối quan hệ như vậy. Nếu để ý kĩ, ta còn thấy giá nhà tang theo lũy thừa nhưng không mạnh (not dramatic).

2. Biểu đồ mối liên hệ độ nóng( Correlation Heatmap)

Trong phần phân tích vừa rồi, ta mới chỉ, dựa theo kinh nghiệm, xét những đặc trưng mà ta nghĩ là có ảnh hưởng tới giá nhà.

Đây là một cách phân tích rất chủ quan và ta cần có một cách phân tích khách quan hơn cho những mối tương quan.

Ta sẽ sử dụng tới Correlation Heatmap. Dưới đây là một heatmap thể hiện mỗi tương quan đôi một giữa những đặc trưng thể hiện được bằng số(continuous features) cũng như mối tương quna của từng đặc trưng này với giá nhà(Sale Price).

(CHO ẢNH Ở ĐÂY Loại mà có 38 feature).

Ý tưởng cơ bản của biểu đồ này là ô vuông càng sáng thì mối tương quan giữa những feature là tọa độ của ô vuông càng mạnh. Dựa vào trực quan này, ta phát hiện thấy một số vùng tối rộng( ví dụ như vùng phía dưới hay bên tay phải của biểu đồ). Điều này có nghĩa là những đặc trưng thể hiện vùng tối này không ảnh hưởng gì nhiều tới những đặc trưng khác. Tất nhiên, một đặc trưng thì hoàn toàn tương quan( completely correlate) với chính nó nên là ta có thể thấy đường chéo của biểu đồ là rất sáng. Tiếp đó, ta sẽ loại bớt đi những đặc trưng không quan trọng này( insignificant feature) và vẽ lại một biểu đồ heatmap khác.

(ChoẢNH SỐ 2)

Ta đã sắp xếp những cột và hàng dựa trên độ tương quan với giá nhà(Sales Price) đối với những đặc trưng sau quá trình ‘’lọc’’ đặc trưng trên.

Ta có một vài nhận xét sau:

* Đúng như ở phần trên, mối tương quan giữa giá nhà và diện tích(GrLivArea) hay chất lượng(OverallQual) là cao(sát với 1).
* Một số đặc trưng có mối quan hệ với giá nhà gần bằng nhau và mối tương quan như với nhau rất cao. Nhìn kĩ vào dữ liệu thì ta có thể thấy hiển nhiên. Ví dụ như đặc trưng diện tích nhà và diện tích sàn tầng 1. Những đặc trưng này có thể coi như gần giống nhau( cùng thể loại với nhau) và ta có thể lọc tiếp để tiếp tục phân tích dữ liệu.

Sau quá trình lọc dữ liệu, ta còn lại 7 đặc trưng đáng quan tâm nhất. Do số đặc trưng là ít, ta vẽ tất cả các biểu đồ (scatterplot) giữa đôi một những đặc trưng.

(ẢNH CÓ NHIỀU scatterplot)

Thông qua những biểu đồ này, ta nhận ra một số mối tương quan thú vị:

* Một số những điểm dữ liệu trong biểu đồ mối quan hệ giữa Diện tích tầng hầm('TotalBsmtSF') và diện tích nhà trên mặt đất(GrLiveArea) tạo thành những đường thẳng(tưởng tượng như những đường biên giới). Điều này được giải thích rằng diện tích hầm sẽ thường không vượt quá diện tích của nhà.
* Một phát hiện thú vị khác dựa vào biểu đồ ở góc dưới cùng phía tay trái. Ta có thể thấy rằng nhà càng mới thì giá càng cao. Ngoài ra, hình dạng tương đối của biểu đồ gần với một hàm lũy thừa nhẹ( hãy vận dụng trí tưởng tượng ở đây). Do đó, ta rút ra rằng giá nhà càng ngaỳ càng có xu hướng tăng nhanh.