TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO

ĐỒ ÁN MÔN HỌC: TRỰC QUAN HÓA DỮ LIỆU

Giảng viên : TS. BÙI TIẾN LÊN

Sinh viên thực hiện :

20120037- Trần Thị Minh Anh
20120125- Bùi Anh Kiệt
20120128 - Nguyễn Thị Cẩm Lai
20120232 - Võ Duy Trường
20120547 - Võ Thành Phong

THÁNG 7/2023

MỤC LỤC

A. THÔNG TIN CHUNG

I. Thông tin thành viên

Họ và tên	MSSV	Mail
Trần Thị Minh Anh	20120037	20120037@student.hcmus.edu.vn
Bùi Anh Kiệt	20120125	20120125@student.hcmus.edu.vn
Nguyễn Thị Cẩm Lai	20120128	20120128@student.hcmus.edu.vn
Võ Duy Trường	20120232	20120232@student.hcmus.edu.vn
Võ Thành Phong	20120547	20120547@student.hcmus.edu.vn

II. Phân chia công việc

Phần	Nhiệm vụ	Người thực hiện
Phần trực quan (60%)	• Lên ý tưởng, thiết kế nội dung cho 1 dashboard/report • Sử dụng bất kỳ công cụ nào; ví dụ, PowerBI, Excel, Tableau, Streamlit, ... để làm 1 bảng dashboard/report trực quan hóa cho bộ dữ liệu	• Minh Anh • Duy Trường • Anh Kiệt
Phần phân tích (40%)	Thống kê mô tả	Cẩm Lai
	Phân tích dữ liệu đơn giản	
	Phân tích hồi quy, giải thích và dự đoán	Thành Phong
	Rút ra nhận xét, thực trạng hiện tại Đề xuất giải pháp	• Thành Phong • Cẩm Lai • Minh Anh • Duy Trường • Anh Kiệt

Viết báo cáo + tổng hợp	Viết báo cáo các nội dung: • Thông tin chung, đánh giá kết quả,... • Bài toán/bộ dữ liệu,… • Phần phân tích dữ liệu	• Cẩm Lai • Thành Phong

III. Đánh giá mức độ hoàn thành

Phần	Chi tiết	Mức độ hoàn thành
Phần trực quan (60%)	• Sử dụng bất kỳ công cụ nào; ví dụ, PowerBI, Excel, Tableau, Streamlit • Triển khai thành công cụ hoàn chỉnh	100%
Phần phân tích: 40%	Thống kê mô tả: 10%	100%
	Phân tích dữ liệu đơn giản: 10%	100%
	Phân tích hồi quy, giải thích và dự đoán: 10%	100%
	Đề xuất giải pháp: 10%	100%
Tổng		100%

IV. Giới thiệu bộ dữ liệu

1. Lý do lựa chọn dữ liệu

- Tính thực tế: Dữ liệu về giá nhà ở Hà Nội là một bộ dữ liệu thực tế và có giá trị, do đó, kết quả phân tích dữ liệu từ dự án này sẽ mang tính ứng dụng cao và có thể được áp dụng trong thực tế.
- Sự đa dạng: Hà Nội là một thành phố lớn với nhiều khu vực khác nhau và các loại hình bất động sản đa dạng. Bộ dữ liệu về giá nhà ở Hà Nội sẽ cung cấp thông tin về

các khu vực, loại hình bất động sản và các yếu tố khác nhau ảnh hưởng đến giá nhà, giúp bạn có cái nhìn tổng quan về thị trường bất động sản của thành phố.

- Tiềm năng kinh doanh: Bất động sản là một lĩnh vực kinh doanh tiềm năng. Bộ dữ liệu về giá nhà ở Hà Nội có thể giúp bạn hiểu rõ hơn về thị trường bất động sản của thành phố, đưa ra các quyết định kinh doanh thông minh và tối ưu hóa lợi nhuận.

2. Nguồn gốc bộ dữ liệu

Bộ dữ liệu được tìm thấy trên Kaggle:

https://www.kaggle.com/datasets/ladcva/vietnam-housing-dataset-hanoi

3. Cấu trúc bộ dữ liệu

- Bộ dữ liệu gồm có:
  - 82496 dòng
  - 12 cột (thuộc tính)
- Các thuộc tính có ý nghĩa sau:

Stt	Tên thuộc tính	Ý nghĩa
1	Diện tích	Mô tả tổng diện tích của căn nhà (m2)
2	Dài	Mô tả chiều dài căn nhà (m)
3	Rộng	Mô tả chiều rộng căn nhà (m)
4	Số phòng ngủ	Mô tả số phòng ngủ có trong căn nhà
5	Số tầng	Mô tả số tầng của căn nhà
6	Địa chỉ	Mô tả địa chỉ của căn nhà (tại Hà Nội)
7	Quận	Mô tả quận nơi căn nhà tọa lạc (tại Hà Nội)
8	Huyện	Mô tả huyện nơi căn nhà tọa lạc (tại Hà Nội)
9	Loại hình nhà ở	Mô tả loại hình nhà ở của căn nhà: + Nhà ngõ, hẻm + Nhà biệt thự + Nhà mặt phố, mặt tiền + Nhà phố liền kề

10	Giá/m2	Giá một mét vương đất (m2)
11	Giấy tờ pháp lý	Mô tả tình trạng giấy tờ pháp lý: + Đã có sổ + Đang chờ sổ + Khác
12	Ngày	Mô tả thông tin ngày mà căn nhà được đăng bán trên website

B. THỐNG KÊ MÔ TẢ

I. Giới thiệu chung

1. Mô tả thống kê là gì?

- Trong phân tích dữ liệu, thống kê mô tả là quá trình tóm tắt và diễn giải thông tin từ các tập dữ liệu. Nó cung cấp cái nhìn tổng quan về các đặc điểm chính của dữ liệu bằng cách sử dụng các chỉ số thống kê khác nhau.
- Thống kê mô tả thường bao gồm việc tính toán và miêu tả các tham số thống kê như trung bình, phương sai, độ lệch chuẩn, trung vị và phân vị. Qua đó, nó cho phép ta hiểu rõ hơn về trung tâm và biến thiên của dữ liệu.
- Ngoài ra, thống kê mô tả cũng thường liên quan đến việc biểu diễn dữ liệu qua các biểu đồ và biểu đồ phân phối như biểu đồ cột, biểu đồ đường, biểu đồ hộp và râu, histogram, đồ thị quantile-quantile (Q-Q plot), và biểu đồ tần số.
- Tổng quan, thống kê mô tả giúp ta có cái nhìn sơ bộ về dữ liệu, hiểu rõ hơn về sự phân bố và tính chất của chúng, từ đó tạo nền tảng cho các phân tích và quyết định tiếp theo.

2. Phương pháp tiếp cận

Nhận thấy thống kê mô tả là bước đi quan trọng cho ta cái nhìn sơ bộ về dữ liệu, do đó nhóm đã thực hiện các bước sau để tiếp cận với thống kê mô tả của bộ dữ liệu:

- Bước 1: Đọc dữ liệu, và khám phá một số đặc điểm cơ bản của tập dữ liệu:
  - Số dòng, số cột của tập dữ liệu
  - Tên các thuộc tính
  - Kiểu dữ liệu các thuộc tính

- Kiểm tra số lượng các giá trị trùng lặp
- Kiểm tra tỉ lệ giá trị null

- Bước 2: Thực hiện tiền xử lý dữ liệu dựa trên các vấn đề phát sinh (cần thiết cho bước thống kê mô tả)
  - Thay đổi kiểu dữ liệu phù hợp cho các cột
    - Từ object sang dạng datetime với thuộc tính "ngày"
    - Từ object sang dạng float với thuộc tính: Số tầng, Số phòng ngủ, Gía/m2, Dài, Rộng
- Bước 3: Tính toán các giá trị thống kê mô tả
  - Tìm ra các giá trị thống kê mô tả cho các thuộc tính như: max, min, mean, median, std, lower quartile (25%), upper quartile (75%), missing ratio
- Bước 4: Sử dụng biểu đồ như biểu đồ cột, biểu đồ tròn, biểu đồ hình thang, biểu đồ violin, và biểu đồ hộp-whisker để trực quan hóa thông tin về phân phối và tập trung của dữ liệu.

II. Kết quả thống kê mô tả

1. Thống kê mô tả của từng cột

	Số tầng	Số phòng ngủ	Diện tích (m2)	Dài (m)	Rộng (m)	Giá/m2 (triệu)
count	36399.000000	82458.000000	82495.000000	19827.000000	35445.000000	8.248400e+04
mean	4.463062	3.881976	51.364438	89.759940	37.814743	9.811330e+03
std	1.573310	1.498314	470.675198	6468.978073	3101.428641	2.546428e+06
min	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000e+00
25%	4.000000	3.000000	34.000000	8.000000	4.000000	7.333000e+01
50%	5.000000	4.000000	40.000000	10.000000	4.000000	9.000000e+01
75%	5.000000	4.000000	50.000000	12.000000	5.000000	1.111100e+02
max	73.000000	11.000000	111411.000000	900000.000000	423432.000000	7.280000e+08

2. Xem xét sự phân bố giá trị của các cột dữ liệu dạng số

titles	Giá/m2 (triệu)	Số tầng	Số phòng ngủ	Diện tích (m2)	Dài (m)	Rộng (m)
missing_ratio	1.500000e-02	55.878	0.046	0.001	75.966	57.034
min	1.000000e+00	1.000	1.000	1.000	1.000	1.000
lower_quartile	7.333000e+01	4.000	3.000	34.000	8.000	4.000
median	9.000000e+01	5.000	4.000	40.000	10.000	4.000
upper_quartile	1.111100e+02	5.000	4.000	50.000	12.000	5.000
max	7.280000e+08	73.000	11.000	111411.000	900000.000	423432.000

3. Xem xét sự phân bố giá trị của các cột dữ liệu không phải dạng số

titles	Quận	Loại hình nhà ở	Giấy tờ pháp lý
missing_ratio	0.001	0.038	35.015
num_values	29	4	3
value_ratios	{'Quận Đống Đa': 16.96, 'Quận Thanh Xuân': 15.709, 'Quận Hoàng Mai': 13.534, 'Quận Hai Bà Trưng': 12.823, 'Quận Hà Đông': 9.495, 'Quận Cầu Giấy': 8.009, 'Quận Ba Đình': 5.688, 'Quận Long Biên': 5.112, 'Quận Nam Từ Liêm': 3.864, 'Quận Tây Hồ': 3.576, 'Quận Bắc Từ Liêm': 1.862, 'Huyện Thanh Trì': 1.456, 'Quận Hoàn Kiếm': 0.64, 'Huyện Hoài Đức': 0.55, 'Huyện Gia Lâm': 0.329, 'Huyện Đông Anh': 0.118, 'Huyện Thanh Oai': 0.062, 'Huyện Sóc Sơn': 0.048, 'Huyện Quốc Oai': 0.029, 'Huyện Đan Phượng': 0.028, 'Huyện Chương Mỹ': 0.024, 'Thị xã Sơn Tây': 0.023, 'Huyện Thường Tín': 0.019, 'Huyện Thạch Thất': 0.017, 'Huyện Mê Linh': 0.013, 'Huyện Ba Vì': 0.007, 'Huyện Phúc Thọ': 0.002, 'Huyện Phú Xuyên': 0.001, 'Huyện Mỹ Đức': 0.001}	{'Nhà ngõ, hẻm': 75.835, 'Nhà mặt phố, mặt tiền': 20.73, 'Nhà phố liền kề': 2.281, 'Nhà biệt thự': 1.154}	{'Đã có sổ': 98.702, 'Đang chờ sổ': 0.664, 'Giấy tờ khác': 0.634}

C. MÔ HÌNH HỌC MÁY

LƯU Ý: Phần code cho các bước được trình bày trong phần báo cáo này được thực hiện trong file notebook đính kèm trong phần nộp bài của nhóm.

I. Bài toán đặt ra

1. Bài toán

Dự đoán giá bán nhà ở, chung cư tại thủ đô Hà Nội, Việt Nam?

2. Giới thiệu chung

- Trong học máy, **học có giám sát** là một nhóm các thuật toán phổ biến trong lĩnh vực này và một trong những vấn đề quan trọng của học có giám sát là hồi quy(regression). Hồi quy là các bài toán liên quan đến việc dự đoán đầu ra có giá trị liên tục (predicting continous valued output).

- Và trong bài toán mà nhóm đề ra thì từ những cột thuộc tính đầu vào như diện tích căn hộ, số tầng, số phòng, ... Nhóm tiến hành dự đoán cột mục tiêu là giá bán (trên m2) của chung cư, nhà ở bằng thuật toán hồi quy tuyến tính (linear regression).

II. Tiền xử lý dữ liệu

1. Loại những thuộc tính không có ý nghĩa cho việc mô hình hóa

- Khi nhìn vào bộ dữ liệu ta có thể thấy ngay có những cột không có ý nghĩa cho việc mô hình hóa khi chúng hoàn toàn không có mối liên quan đến biến mục tiêu. Loại bỏ sớm các thuộc tính như vậy sẽ giúp việc tiền xử lý dữ liệu được dễ dàng hơn.

- Ta có thể thấy:

+ Cột ‘Ngày’, ‘Địa chỉ’ và ‘Huyện’ **chứa các giá trị quá riêng biệt**, không có ý nghĩa cho việc trực quan hay phân tích, chúng được đăng để cung cấp đầu đủ thông tin rõ

ràng cho một căn nhà cần được bán, hoặc theo format của trang web dùng để đăng bán, do đó các cột này sẽ không được lựa chọn làm thuộc tính đầu vào.

+ Thuộc tính 'Diện tích' được tính bằng công thức 'Dài'*'Rộng', bên cạnh đó có nhiều mẫu dữ liệu không có thông tin về chiều dài và chiều rộng căn nhà mà chỉ có tổng diện tích, do đó có thể loại bỏ hai thuộc tính này để giảm độ phức tạp của mô hình.

+ Các cột của bộ dữ liệu được giữ lại bao gồm: **'Quận', 'Loại hình nhà ở', 'Giấy tờ pháp lý', 'Số tầng', 'Số phòng', 'Diện tích', 'Giá/m2'**.

- Đây là bước loại những thuộc tính có thể thấy ngay về mặt ý nghĩa, sau khi hoàn thành các bước tiền xử lý tiếp theo sẽ thực hiện tính correlation (hệ số tương quan) giữa từng thuộc tính với biến đầu ra để có thể đưa ra những lựa chọn chính xác hơn nữa.

2. Chuyển đổ các cột không phải dạng số về dạng số

- Dữ liệu đầu vào cho các mô hình học máy đều phải ở dang số, do đó cần chuyển đổi các cột không phải dạng số về dạng số

- Có nhiều cách để biến đổi cột dạng danh mục về dạng số như one-hot vector, ordinal label, sử dụng một giá trị số làm đại diện, hoặc đơn giản là thay đổi trực tiếp kiểu dữ liệu của cột dự liệu đó. Việc chuyển đổi của nhóm được thực hiện như sau:

+ Nhận thấy tất cả các thuộc tính dạng danh mục của bộ dữ liệu đều có nhiều hơn 2 loại giá trị. Đối với các thuộc tính có nhiều hơn 2 loại giá trị, chúng ta sẽ má hõa bằng one-hot vector, lý do sử dụng one-hot mà không dùng ordinal hay label là để **tránh xảy ra hiện tượng bias do mã hóa thành các giá trị lớn nếu số lượng các giá trị trong cột dữ liệu là lớn.** Ví dụ đơn giản như cột 'Quận', có hơn 30 quận ở Hà Nội và có những Quận mà mức giá đất của chúng ngang nhau. Khi sử dụng ordinal hay label để chuyển hóa thành dạng số, các quận sẽ được đánh số từ 1 đến hơn 30, khi đó hiện tượng bias sẽ xảy ra trong trường hợp ví dụ như quận thứ 1 và quận thứ 30 có mức ảnh hưởng đến giá đất như nhau nhưng những mẫu thuộc quận thứ 30 sẽ bị phạt nặng hơn vì giá trị lên đến 30.

+ Ngoài ra các giá trị của các cột 'Số tầng', 'Số phòng ngủ', 'Diện tích', 'Giá/m2' có thể giữ nguyên lại giá trị số của nó mà không cần dùng one-hot vì số tầng hay số phòng

ngủ càng nhiều thì càng ảnh hưởng đến giá của căn nhà. 'Diện tích' và 'Giá/m2' mang kiểu categorical vì chỉ có thêm phần đơn vị.

3. Tính Correlations và tiếp tục chọn những thuộc tính thật sự có giá trị

- Correlation là một thuật ngữ thống kê được sử dụng phổ biến đề cập đến mức độ liên quan của hai biến để có mối quan hệ tuyến tính với nhau hay không.

- Correlation cao nhất có giá trị là 1 (hai biến hoàn toàn có quan hệ tuyến tính) và thấp dần nếu hai biến càng không có quan hệ tuyến tính.

- Nhóm sẽ chọn ra những thuộc tính hoàn toàn không có liên quan gì đến 'Giá/m2' bằng cách loại các cột cho ra điểm correlation với 'Giá/m2' là không.

	Số tầng	Số phòng ngủ	Diện tích	Giá/m2	Quận_Huyện Ba Vì	Quận_Huyện Chương Mỹ	Quận_Huyện Gia Lâm	Quận_Huyện Hoài Đức	Quận_Huyện Mê Linh	Quận_Huy Mỹ E
Số tầng	1.000000	0.375766	-0.000289	-0.011569	NaN	-0.013301	-0.044457	-0.047928	-0.011415	N
Số phòng ngủ	0.375766	1.000000	0.018516	-0.006874	0.000672	-0.014367	-0.003113	-0.031267	-0.001893	0.002
Diện tích	-0.000289	0.018516	1.000000	0.000098	0.027463	0.004881	0.005836	0.003166	0.003836	0.000
Giá/m2	-0.011569	-0.006874	0.000098	1.000000	-0.000033	-0.000060	-0.000220	-0.000284	-0.000044	-0.000
Quận_Huyện Ba Vì	NaN	0.000672	0.027463	-0.000033	1.000000	-0.000133	-0.000490	-0.000634	-0.000098	-0.000
Quận_Huyện Chương Mỹ	-0.013301	-0.014367	0.004881	-0.000060	-0.000133	1.000000	-0.000894	-0.001158	-0.000180	-0.000
Quận_Huyện Gia Lâm	-0.044457	-0.003113	0.005836	-0.000220	-0.000490	-0.000894	1.000000	-0.004271	-0.000663	-0.000
Quận_Huyện Hoài Đức	-0.047928	-0.031267	0.003166	-0.000284	-0.000634	-0.001158	-0.004271	1.000000	-0.000859	-0.000

- Correlations của 'Giá/m2' với hầu hết các cột khác đều rất nhỏ nhưng không bằng 0 chứng tỏ đây là dấu hiệu cho thấy mực độ phân tán cực kỳ lớn của bộ dữ liệu này, có thể mô hình hồi quy sẽ gặp khó khăn.

4. Xử lý các giá trị NaN

- Xóa đi những mẫu có cột 'Giá/m2' là Nan.

- Đối với bộ dữ liệu này mỗi thuộc tính có số lượng giá trị thiếu khá nhiều do đó việc bỏ đi các dòng chứa giá trị nan có thể gây ảnh hưởng lớn đến tính chính xác khi tiến hành học trên bộ dữ liệu do thiếu thông tin.

- Giải pháp có thể sử dụng là thay thế giá trị NaN bằng các giá trị đặc biệt của cột dữ liệu chẳng hạn: trung bình, trung vị, most, ...

- Nhóm sẽ sử dụng giá trị median để thay thế các giá trị NaN cho thuộc tính 'Diện tích', do dữ liệu về thuộc tính diện tích có thể có tồn tại điểm ngoại lệ có giá trị quá lớn nên việc dùng giá trị mean có thể gây ra sai sót. Và dùng giá trị most_frequent để thay thế các giá trị NaN cho các thuộc tính định lượng rời rạc còn lại (trừ thuộc tính 'Số tầng').

- Đặc biệt với thuộc tính 'Số tầng' do có quá nhiều mẫu bị thiếu - đến gần phân nửa số mẫu bị thiếu thuộc tính 'Số tầng' do đó không đủ thông tin để thực hiện chọn chiến lược điền giá trị thiếu, do đó nhóm sẽ tạo một giá trị mới là -1 để giữ nguyên tính toàn vẹn của dữ liệu.

III. Xây dựng mô hình học máy

1. Hồi quy tuyến tính đa biến (Linear Regression with Multi Variables)

a) Mô tả bài toán

- Hồi quy tuyến tính đa biến áp dụng nhiều biến thuộc tính để xây dựng hypothesis cho mô hình hồi quy.

- Hypothesis chỉ có thể được xây dựng gần với giá trị thật của biến đầu ra nhất khi thuộc tính được chọn phải có mối quan hệ tuyến tính thật sự với biến y và quan trọng hơn hết nó có vai trò ảnh hưởng phải lớn đến giá trị của biến y. Đối vối bài toán dự đoán giá bán của căn hộ/chung cư thì có thể chấp nhận việc thuộc tính diện tích ngôi nhà là có vai trò ảnh hưởng lớn đến giá nhà nhưng vẫn có khả năng số phòng ngủ hay số phòng vệ sinh sẽ cũng ảnh hưởng nhiều đến giá bán.

b) Regularization

- Khi dùng càng nhiều biến thì độ phức tạp của hypothesis càng tăng lên, dẫn đến có thể tạo ra những dạng đường ngoằn ngòe quá mức cần thiết và không thực tế, gọi là vấn đề over fitting.

- Có nhiều cách để khắc phục vấn đề over fitting này như là các thuật toán lựa chọn mô hình (Model Selection) hay Regularization.

- Trong đồ án lần này nhóm sẽ sử dụng phương pháp Regularization.

c) Cost Function

- Ý tưởng thực hiện regularization đó chính là thêm vào cost function một đại lượng nữa là tổng các tích của một hằng số λ với vector tham số θ của hypothesis.

- Đại lượng này tác động đến Cost Function như sau:

$$J(\theta) = \frac{1}{2m}[\sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 + \lambda \sum_{j=1}^{n} \left(\theta_j\right)^2]$$

  - Cùng nhớ lại mục tiêu của hồi quy tuyến tính là minimize hàm Cost Function do đó khi thêm một đại lượng là $\lambda \sum_{j=1}^{n} \left(\theta_j\right)^2$ thì để cho có thể minimize được Cost Function bắt buộc các tham số θ_j phải nhỏ và ta gọi việc 'ép buộc' này là phạt (penalize) các tham số θ.
  - Tuy nhiên nhìn vào công thức của đại lượng được thêm vào bạn chắc chắn sẽ chú ý đến hằng số λ, vậy việc chọn λ ảnh hưởng như nào đến quá trình Regularization.
  - Nếu chúng ta thiết lập λ rất nhỏ thì việc thực hiện regularization sẽ không còn ý nghĩa nữa, còn nếu như thiết lập λ rất lớn thì việc penalize các tham số θ là rất nặng dẫn đến θ_j (j=[1, n]) sẽ xấp xỉ 0 và lúc này hypothesis trở thành một hàm hằng $h_\theta(x) = \theta_0$ dẫn đến hiện tượng Underfitting.
  - Để dễ dàng hơn nhóm sẽ chọn λ=1 cho quá trình này.

d) Gradient Descent

- các tham số θ sẽ là những giá trị mà chúng ta cần phải thay đổi để tối ưu hóa Cost Function, và một trong những cách để thực hiện việc này là thuật toán Gradient Descent.

- Thuật toán được thực hiện như sau:

Trong mỗi lần lặp cập nhật một cách đồng thời các tham số θ_j theo công thức như sau:

$$\theta_0 = \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right) x_0^{(i)}$$

$$\theta_j = \theta_j(1 - \alpha \frac{\lambda}{m}) - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right) x_j^{(i)} \text{ với j=[1, n] nếu x có n thuộc tính.}$$

Trong đó: alpha là 'learning rate' giúp việc học được tối ưu hơn.

Normal Equation

- Ngoài việc đạo hàm để cập nhật tham số, đối với các bộ dữ liệu lớn để tiết kiệm thời gian chúng ta có thể sử dụng một cách khác để tìm nghiệm θ tối ưu nhất đó chính là 'Normal Equation'.

$$\theta = \left(X^T X + \lambda A\right)^{-1} X^T \vec{y}$$

- Với A là ma trận chứa các giá trị 0 và chỉ có các giá trị trên đường chéo chính có giá trị là 1 nhưng trừ phần tử ở vị trí [0, 0] cũng bằng 0.

2. Cài đặt

Các bước cài đặt được trình bày chi tiết trong file notebook source code mà nhóm nộp kèm theo.

- **Đánh giá mô hình:**

- Nhóm sử dụng độ đo R^2 score để đánh giá hiệu suất mô hình hồi quy.

R^2 được định nghĩa bằng công thức:

$$(1 - \frac{u}{v})$$

với u = sum((y_true - y_pred)^2)

v = sum((y_true - y_true.mean( ))^2)

- Độ đo này đánh giá độ phù hợp của mô hình. Điểm số tốt nhất có thể có là 1.0 và có thể có giá trị âm (mô hình cho ra kết quả quá tệ). Nó sẽ cho biết tỷ lệ các điểm dữ liệu nằm gần đường tuyến tính như thế nào. Nếu càng nhiều điểm dữ liệu trong bộ dữ liệu nằm gần đường tuyến tình thì điểm càng cao và ngược lại.

- R^2 score thu được cho bộ dữ liệu này khi áp dụng hồi quy tuyến tính là 0.15, bên cạnh đó giá trị Cost trên tập Train là 283.5 và trên tập Validation là 288.6.

- Normal Equation cho kết quả tốt hơn đôi chút khi Cost trên tập Validation là 249.3.

- NHẬN XÉT:

+ R^2 score chỉ đạt 0.15, từ đó có thể thấy được tính không phù hợp của mô hình hồi quy tuyến tính trên bộ dữ liệu này.

+ Lý do không phù hợp có thể là:

* Nếu chỉ xét riêng về vị trí địa lý thì những quận ở trung tâm như Cầu Giấy, Hoàn Kiếm, Ba Đình, Hai Bà Trưng, Đống Đa sẽ có giá đất rất mắc hơn so với các quận vùng ven khác như Long Biên, Thanh Trì.

* Tuy nhiên ở các vùng trung tâm thì mật độ dân số lại đông và diện tích xây dựng được phép sẽ hẹp hơn, ta có thể thấy diện tích ngôi nhà thường lớn ở những quận không phải trung tâm trong bộ dữ liệu cũng sẽ ảnh hưởng về giá.

* Mô hình có thể dự đoán giá của các căn hộ có giá bán từ thấp đến trung bình, nhưng không dự đoán tốt với các căn hộ có giá bán cao.

* Còn nhiều yếu tố ngoại cảnh ảnh hưởng đến giá bán như chất lượng xây dựng, chất lượng nội thất, giá thầu, ...

* Có thể nhiều chủ căn hộ đăng giá theo cảm tính, như ở phần xem xét sự phân bố của một số thuộc tính thì khi nhìn vào bộ dữ liệu có rất nhiều điểm dữ liệu xếp chồng lên nhau, điều này chứng tỏ với những kích thước gần nhau nhưng giá rất khác biệt nhau.

IV. Đề xuất giải pháp

- Sự không phù hợp của hồi quy tuyến tính có thể được khắc phục bằng một số đề xuất như sau:

+ Hồi quy đa thức (Polynomial Regression). Tuy nhiên có thể giải pháp này có thể cũng chưa tốt hơn nhiều so với hồi quy tuyến tính vì sự phân bố của dữ liệu có độ phân tán cực kỳ lớn, không có hình dạng của một hàm nhất định nào.

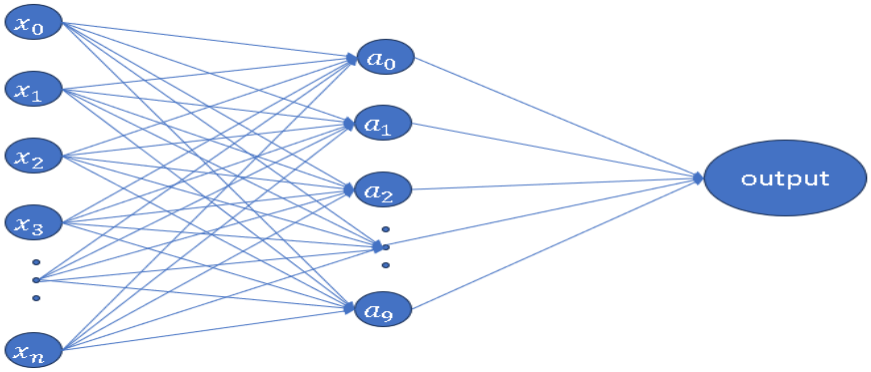
+ Random Forest với những cây con là Regressor Tree.

+ Sử dụng một mạng neural nông (Shallow Neural Network) để phân tích hồi quy.

- Và trong bài làm của nhóm, nhóm sẽ thực hiện giải pháp Shallow Neural Network cho việc cải tiến quy trình phân tích hồi quy.

V. Shallow neural network

- Kiến trúc mạng:



- Shallow Neural Network là một mạng neural nông gồm một lớp đầu vào, một lớp ẩn và một lớp đầu ra như hình trên. Lớp ẩn nhóm chúng em sẽ sử dụng 10 units trong lớp này và activation function được sử dụng là **hàm ReLU**.

- **Phần cài đặt được trình bày chi tiết trong file notebook source code mà nhóm nộp kèm theo.**

- R^2 score dao động trong khoảng 0.3 đến 0.31, một kết quả tốt hơn khá nhiều với mô hình hồi quy truyền thống. Với một bộ dữ liệu phân tán cực lớn nhưng R^2 score vẫn dương và đạt tới trên 0.3 là một thành công cho mô hình mạng học nông.