TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG



**BÁO CÁO HỌC MÁY VÀ**

**KHAI PHÁ DỮ LIỆU**

ĐỀ TÀI : Dự đoán giá nhà

***GVHD* : TS. Thân Quang Khoát**

***NHÓM* : 3**

***SVTH*****: Nguyễn Mạnh Cường 20172989**

**Nguyễn Quang Phúc 20173304**

**Nguyễn Đình Đạt 20173011**

**Trần Thị Thúy 20173394**

***Lớp* : Học máy và Khai phá dữ liệu**  **120446**

***Hà Nội, tháng 12 năm 2020***

# I. Giới thiệu đề tài

## 1. Giới thiệu bài toán

Trong những năm gần đây, với sự phát triển mạnh mẽ của máy tính và dữ liệu, lĩnh vực Trí tuệ nhân tạo đã có nhiều thành tựu và được áp dụng rộng rãi. Đặc biệt, các giải thuật học máy đã và đang giải quyết rất tốt một số bài toán thực tế, bởi tính phù hợp cao.

Bài toán dự đoán giá bất động sản hay cụ thể hơn là nhà ở, rất được quan tâm. Giá trị của một ngôi nhà, một mảnh đất hay một tài sản bất động sản nói chung thường được con người quyết định bởi nhiều yếu tố khác nhau như diện tích, số phòng, số tầng, khu vực… đến các yếu tố phức tạp hơn như tình hình kinh tế, chính trị, xã hội trong nước, vấn đề tâm linh, …

Chính vì vậy, để áp dụng những kiến thức đã học được trong học phần Nhập môn Học máy và Khai phá dữ liệu, nhóm chúng em đã thực hiện giải quyết bài toán : Dự đoán giá nhà.

## 2. Mô tả bài toán

Input: Tập dữ liệu chứa các thuộc tính của các ngôi nhà và giá cả

Output: Tập giá nhà được dự đoán với độ chính xác cao.

# II. Cơ sở lý thuyết

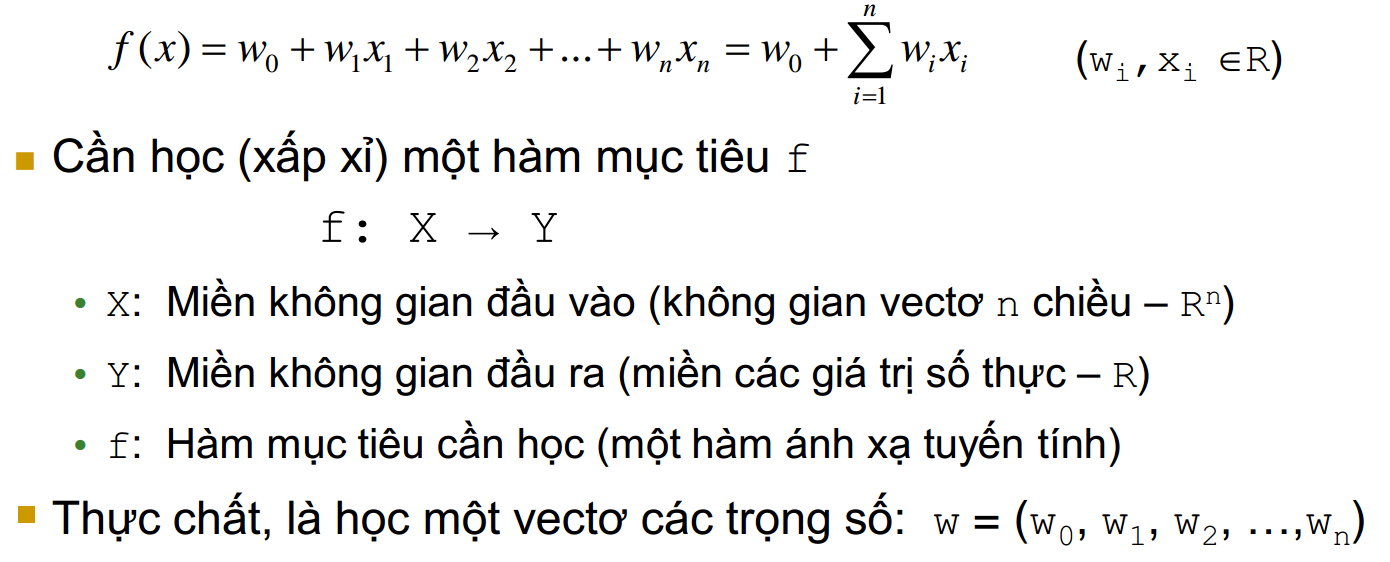
## 1. Regression

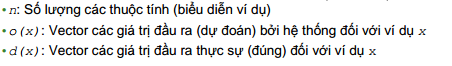
Regression là phương pháp nghiên cứu mối quan hệ giữa 2 biến mà cụ thể một biến sẽ là độc lập (ảnh hưởng đến biến mich tiêu) và biến còn lại sẽ là biến mục tiêu (bị ảnh hưởng bởi biến độc lập), nó mô hình hóa, định lượng hóa mối quan hệ này để qua đó có thể xác định được giá trị của biến mục tiêu nếu các biến độc lập thay đổi như thế nào. Regression là phương pháp chính trong Predictive analytics (phân tích dự báo) bên cạnh là kiến thức nền tảng trong lịch vực thống kế (Statistics) và khai phá dữ liệu (Data mining).

### 1.1. Linear Regression

Linear Regression là mô hình ước lượng mối quan hệ giữa các thuộc tính một cách và dự đoán thuộc tính có miền giá trị liên tục thông qua một hàm tuyến tính.

Hàm mục tiêu là một hàm tuyến tính :





Để giải quyết bài toán này có thể sử dụng phương pháp bình phương tối thiểu (OLS):

Input: D :

Output: w\* - bộ trọng số tối ưu

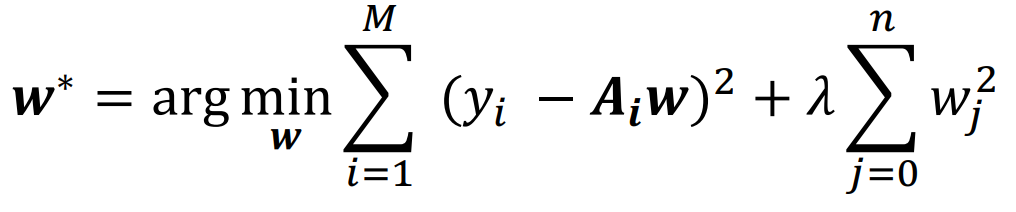
Học:

A là ma trận tạo bởi các điểm dữ liệu

### 1.2. Ridge Regression

Phương pháp OLS có nhược điểm là khối lượng tính toán lớn, cũng như dễ bị quá khớp và không áp dụng được nếu ma trận không khả nghịch.

Để khắc phục các nhược điểm này, chúng ta có thể thêm các thành phần chính quy hóa (Regularization) vào hàm mục tiêu. Nếu thêm một thành phần chính quy hóa là L2 norm thì ta gọi là Ridge Regression .Khi đó bài toán có dạng:

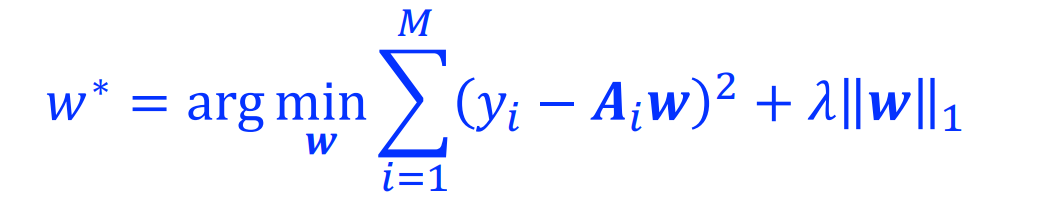


là hằng số chính quy hóa. là ma trận tạo bởi các điểm dữ liệu

Phương pháp này giúp hạn chế quá khớp và tránh được trường hợp ma trận không khả nghịch như OLS, tuy nhiên nhược điểm là lỗi có thể cao hơn OLS. Ngoài ra khả năng dự đoán của phương pháp này cũng phụ thuộc đáng kể vào cách chọn siêu tham số

### 1.2 Lasso Regression

Tương tự Ridge Regression, Lasso Regression thực chất là thêm một thành phần chính quy hóa L1-norm:



Mục đích của việc thêm thành phần chính quy hóa L1-norm để giảm thiểu các đặc trưng không quan trọng , từ đó đưa ra những lời giải thưa hơn

## 2. Random forest

Cây quyết định là một cây phân cấp có cấu trúc dùng để phân lớp các đối tượng dựa vào dãy các luật. Mỗi luật của cây được diễn giải theo dạng: nếu…. thì…. Cây quyết định được sử dụng trong rất nhiều các ứng dụng thực tế và cuộc sống.

Cấu trúc của một cây quyết định gồm 3 phần:

* một nút gốc và các nút con chứa thuộc tính cần được kiểm tra của dữ liệu đầu vào.
* các nút lá: biểu diễn các nhãn cần dự đoán.
* các nhánh của một nút: biểu diễn cho giá trị của thuộc tính tại nút đó.

Sử dụng cây quyết định để dự đoán nhãn cho một đối tượng đầu vào bằng cách sử dụng các thuộc tính của đối tượng đó để đi chuyển từ gốc đến một lá. Nhãn nằm trong lá là nhãn cần sự đoán.

Random forest là một mô hình học có giám sát, hoạt động dựa trên việc kết hợp nhiều cây quyết định tạo ra một rừng cây.

Cách thức hoạt động của thuật toán:

* Chọn ra ngẫu nhiên k features từ tập m features (k<<m)
* Xây dựng cây quyết định dựa trên k features đã chọn
* Lặp lại 2 bước trên để tạo ra n cây.
* Trong bài toán hồi quy, với một bản ghi mới, mỗi cây sẽ dự đoán ra một giá trị y. kết quả cuối cùng được tính bằng giá trị trung bình của tất cả các dự đoán của tất cả các cây trong rừng. trong bài toán phân loại, nỗi cây sẽ dự đoán ra một nhãn cho bản ghi đầu vào, sau đó chọn ra nhãn được dự đoán nhiều nhất làm nhãn cho bản ghi đó.

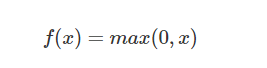
## 3. Artificial neural network

Cấu trúc ANN gồm 3 thành phần chính: Input layer cà output layer chỉ gồm 1 layer, hidden layer có thể có 1 hay nhiều layer ( với bài toán này gồm 3 tầng hidden layer). ANN hoạt động theo hướng mô tả lại cách hoạt theo hướng mô tả lại cách hoạt động của hệ thần kinh với neuron được kết nối với nhau.

Trong ANN, trừ input layer thì tất cả các node thuộc các layer khác đều full-connected với các node thuộc layer trước nó. Mỗi node thuộc hidden layer nhận vào ma trận đều vào từ layer trước và kết hợp với trọng số để ra được kết quả.

Hàm kích hoạt (activation function) mô phỏng tỷ lệ truyền xung qua axon của một neuron thần kinh.Trong bài toán này nhóm bọn em dùng hàm ReLU

Công thức

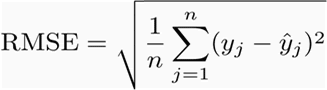


Cấu trúc mạng cụ thể:

* Lớp 1: Lớp kết nối đầy đủ (Dense), nhận đầu vào là một vector 174 đầu ra là 1 vector kích thước 50 , hàm kích hoạt ReLU
* Lớp 2: Lớp kết nối đầy đủ (Dense), nhận đầu vào là một vector 50 đầu ra là 1 vector kích thước 25 , hàm kích hoạt ReLU
* Lớp 3: Lớp kết nối đầy đủ (Dense), nhận đầu vào là một vector 25 đầu ra là 1 vector kích thước 50, hàm kích hoạt ReLU
* Lớp 4: Lớp kết nối đầy đủ (Dense), nhận đầu vào là một vector 50 đầu ra là kết quả cần dự đoán

Hàm ReLU đang được sử dụng khá nhiều trong những năm gần đây khi huấn luyện các mạng neuron. ReLU đơn giản lọc các giá trị < 0. Nó có nhiều ưu điểm về tốc độ, tính toán nhanh nhưng với các node có giá trị nhỏ hơn 0, qua ReLU activation sẽ thành 0, hiện tượng đấy gọi là “Dying ReLU” Nếu các node bị chuyển thành 0 thì sẽ không có ý nghĩa với bước linear activation ở lớp tiếp theo.

Hàm loss nhóm bọn em sử dụng: root\_mean\_squared\_error với công thức



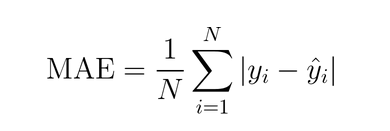
Khởi tạo trọng số có sẵn (init = “he\_uniform”) : nhóm bọn em sử dụng lấy mẫu từ một phân phối đồng đều trong [-limit, limit], trong đó limit = sqrt (6 / fan\_in) (fan\_in là số lượng đơn vị đầu vào trong tensor trọng lượng).

## 4. Đánh giá

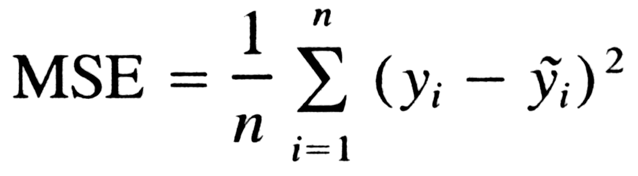
Nhóm sử dụng các thang đo:

Accuracy: Độ chính xác của dự đoán

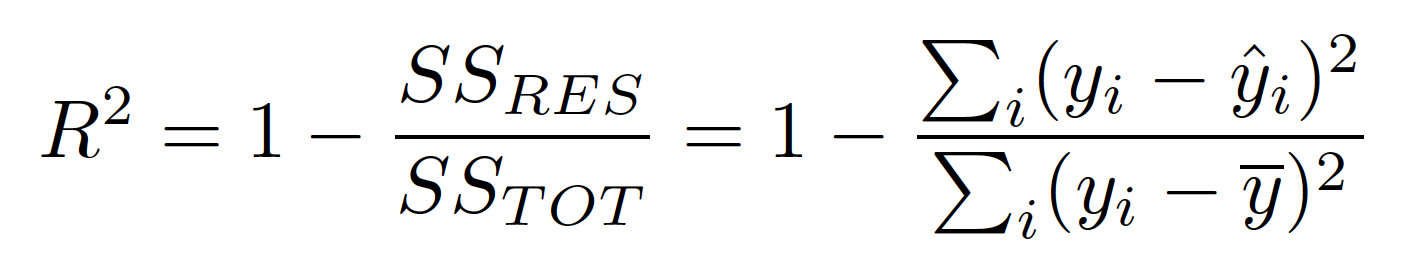
MAE: Đặc trưng cho độ lệch của dự đoán theo norm1



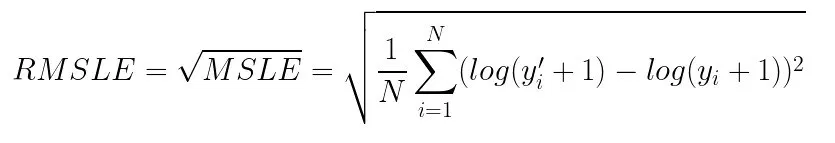
MSE :Đặc trưng cho độ lệch của dự đoán theo norm2



R2-score: đặc trưng cho sự phù hợp của đường hồi quy với dữ liệu, là một số có giá trị từ 0 đến 1, được tính theo công thức



RMSLE:Root Mean Square Log Error:



# III. Dữ liệu

## 1. Thống kê từ dữ liệu gốc theo các thuộc tính

### 1.1. Giá trị nhỏ nhất, lớn nhất của mỗi thuộc tính

#### a) Bộ dữ liệu dùng để train:

Giá trị nhỏ nhất

|  |  |  |
| --- | --- | --- |
| MSSubClass: 20  MSZoning: C (all)  LotFrontage: 21.0  LotArea: 1300  Street: Grvl  LotShape: IR1  LandContour: Bnk  Utilities: AllPub  LotConfig: Corner  LandSlope: Gtl  Neighborhood: Blmngtn  Condition1: Artery  Condition2: Artery  BldgType: 1Fam  HouseStyle: 1.5Fin  OverallQual: 1  OverallCond: 1  3SsnPorch: 0  ScreenPorch: 0  PoolArea: 0  SalePrice: 34900 | YearBuilt: 1872  YearRemodAdd: 1950  RoofStyle: Flat  RoofMatl: ClyTile  Exterior1st: AsbShng  Exterior2nd: AsbShng  MasVnrArea: 0.0  ExterQual: Ex  ExterCond: Ex  Foundation: BrkTil  BsmtFinSF1: 0  BsmtFinSF2: 0  BsmtUnfSF: 0  TotalBsmtSF: 0  Heating: Floor  HeatingQC: Ex  CentralAir: N  1stFlrSF: 334  2ndFlrSF: 0  MiscVal: 0  MoSold: 1  YrSold: 2006 | LowQualFinSF: 0  GrLivArea: 334  BsmtFullBath: 0  BsmtHalfBath: 0  FullBath: 0  HalfBath: 0  BedroomAbvGr: 0  KitchenAbvGr: 0  KitchenQual: Ex  TotRmsAbvGrd: 2  Functional: Maj1  Fireplaces: 0  GarageYrBlt: 1900.0  GarageCars: 0  GarageArea: 0  PavedDrive: N  WoodDeckSF: 0  OpenPorchSF: 0  EnclosedPorch: 0  SaleType: COD  SaleCondition: Abnorml |

Giá trị lớn nhất

|  |  |  |
| --- | --- | --- |
| MSSubClass: 190  MSZoning: RM  LotFrontage: 313.0  LotArea: 215245  Street: Pave  LotShape: Reg  LandContour: Lvl  Utilities: NoSeWa  LotConfig: Inside  LandSlope: Sev  Neighborhood: Veenker  Condition1: RRNn  Condition2: RRNn  BldgType: TwnhsE  HouseStyle: SLvl  OverallQual: 10  OverallCond: 9  YearBuilt: 2010  YearRemodAdd: 2010  RoofStyle: Shed  RoofMatl: WdShngl | MasVnrArea: 1600.0  ExterQual: TA  ExterCond: TA  Foundation: Wood  BsmtFinSF1: 5644  BsmtFinSF2: 1474  BsmtUnfSF: 2336  TotalBsmtSF: 6110  Heating: Wall  HeatingQC: TA  CentralAir: Y  1stFlrSF: 4692  2ndFlrSF: 2065  LowQualFinSF: 572  GrLivArea: 5642  BsmtFullBath: 3  BsmtHalfBath: 2  FullBath: 3  HalfBath: 2  BedroomAbvGr: 8  Exterior1st: WdShing  Exterior2nd: Wd Shng | KitchenAbvGr: 3  KitchenQual: TA  TotRmsAbvGrd: 14  Functional: Typ  Fireplaces: 3  GarageYrBlt: 2010.0  GarageCars: 4  GarageArea: 1418  PavedDrive: Y  WoodDeckSF: 857  OpenPorchSF: 547  EnclosedPorch: 552  3SsnPorch: 508  ScreenPorch: 480  PoolArea: 738  MiscVal: 15500  MoSold: 12  YrSold: 2010  SaleType: WD  SaleCondition: Partial  SalePrice: 755000 |

#### b) Bộ dữ liệu dùng để Test:

Giá tri nhỏ nhất

|  |  |  |
| --- | --- | --- |
| MSSubClass: 20  LotFrontage: 21.0  LotArea: 1470  Street: Grvl  LotShape: IR1  LandContour: Bnk  LotConfig: Corner  LandSlope: Gtl  Neighborhood: Blmngtn  Condition1: Artery  Condition2: Artery  BldgType: 1Fam  HouseStyle: 1.5Fin  OverallQual: 1  OverallCond: 1  YearBuilt: 1879  YearRemodAdd: 1950  RoofStyle: Flat  RoofMatl: CompShg | ExterQual: Ex  ExterCond: Ex  Foundation: BrkTil  BsmtFinSF1: 0.0  BsmtFinSF2: 0.0  BsmtUnfSF: 0.0  TotalBsmtSF: 0.0  Heating: GasA  HeatingQC: Ex  CentralAir: N  Electrical: FuseA  1stFlrSF: 407  2ndFlrSF: 0  LowQualFinSF: 0  GrLivArea: 407  BsmtFullBath: 0.0  BsmtHalfBath: 0.0  FullBath: 0  HalfBath: 0 | GarageYrBlt: 1895.0  GarageCars: 0.0  GarageArea: 0.0  PavedDrive: N  WoodDeckSF: 0  OpenPorchSF: 0  EnclosedPorch: 0  3SsnPorch: 0  ScreenPorch: 0  PoolArea: 0  MiscVal: 0  MoSold: 1  YrSold: 2006  SaleCondition: Abnorml  TotRmsAbvGrd: 3  Fireplaces: 0  BedroomAbvGr: 0  KitchenAbvGr: 0  MasVnrArea: 0.0 |

Giá trị lớn nhất

|  |  |  |
| --- | --- | --- |
| MSSubClass: 190  LotFrontage: 200.0  LotArea: 56600  Street: Pave  LotShape: Reg  LandContour: Lvl  LotConfig: Inside  LandSlope: Sev  Neighborhood: Veenker  Condition1: RRNn  Condition2: PosN  BldgType: TwnhsE  HouseStyle: SLvl  OverallQual: 10  OverallCond: 9  YearBuilt: 2010  YearRemodAdd: 2010  RoofStyle: Shed  RoofMatl: WdShngl | MasVnrArea: 1290.0  ExterQual: TA  ExterCond: TA  Foundation: Wood  BsmtFinSF1: 4010.0  BsmtFinSF2: 1526.0  BsmtUnfSF: 2140.0  TotalBsmtSF: 5095.0  Heating: Wall  HeatingQC: TA  CentralAir: Y  Electrical: SBrkr  1stFlrSF: 5095  2ndFlrSF: 1862  LowQualFinSF: 1064  GrLivArea: 5095  BsmtFullBath: 3.0  BsmtHalfBath: 2.0  SaleCondition: Partial | FullBath: 4  HalfBath: 2  BedroomAbvGr: 6  KitchenAbvGr: 2  TotRmsAbvGrd: 15  Fireplaces: 4  GarageYrBlt: 2207.0  GarageCars: 5.0  GarageArea: 1488.0  PavedDrive: Y  WoodDeckSF: 1424  OpenPorchSF: 742  EnclosedPorch: 1012  3SsnPorch: 360  ScreenPorch: 576  PoolArea: 800  MiscVal: 17000  MoSold: 12  YrSold: 2010 |

### 1.2. Giá trị trung bình và độ lệch chuẩn

#### a) Dữ liệu dùng để train

Giá trinh trung bình (mean):

|  |  |  |
| --- | --- | --- |
| MSSubClass 56.897260  LotFrontage 70.049958  LotArea 10516.828082  OverallQual 6.099315  OverallCond 5.575342  YearBuilt 1971.267808  YearRemodAdd 1984.865753  MasVnrArea 103.685262  BsmtFinSF1 443.639726  BsmtFinSF2 46.549315  BsmtUnfSF 567.240411  TotalBsmtSF 1057.429452 | LowQualFinSF 5.844521  GrLivArea 1515.463699  BsmtFullBath 0.425342  BsmtHalfBath 0.057534  FullBath 1.565068  HalfBath 0.382877  BedroomAbvGr 2.866438  KitchenAbvGr 1.046575  TotRmsAbvGrd 6.517808  Fireplaces 0.613014  GarageYrBlt 1978.506164  GarageCars 1.767123  GarageArea 472.980137 | OpenPorchSF 46.660274  EnclosedPorch 21.954110  3SsnPorch 3.409589  ScreenPorch 15.060959  PoolArea 2.758904  MiscVal 43.489041  MoSold 6.321918  YrSold 2007.815753  SalePrice 180921.195890  WoodDeckSF 94.244521  1stFlrSF 1162.626712  2ndFlrSF 346.992466 |

Giá trị độ lệch chuẩn (std):

|  |  |  |
| --- | --- | --- |
| MSSubClass 42.300571  LotFrontage 24.284752  LotArea 9981.264932  OverallQual 1.382997  OverallCond 1.112799  YearBuilt 30.202904  YearRemodAdd 20.645407  MasVnrArea 181.066207  BsmtFinSF1 456.098091  BsmtFinSF2 161.319273  OpenPorchSF 66.256028  EnclosedPorch 61.119149  3SsnPorch 29.317331 | GrLivArea 525.480383  BsmtFullBath 0.518911  BsmtHalfBath 0.238753  FullBath 0.550916  HalfBath 0.502885  BedroomAbvGr 0.815778  KitchenAbvGr 0.220338  TotRmsAbvGrd 1.625393  Fireplaces 0.644666  GarageYrBlt 24.689725  GarageCars 0.747315  GarageArea 213.804841  WoodDeckSF 125.338794 | ScreenPorch 55.757415  PoolArea 40.177307  MiscVal 496.123024  MoSold 2.703626  YrSold 1.328095  SalePrice 79442.502883  BsmtUnfSF 441.866955  TotalBsmtSF 438.705324  1stFlrSF 386.587738  2ndFlrSF 436.528436  LowQualFinSF 48.623081 |

Giá trị trung vị (median)

|  |  |  |
| --- | --- | --- |
| MSSubClass 50.0  LotFrontage 69.0  LotArea 9478.5  OverallQual 6.0  OverallCond 5.0  YearBuilt 1973.0  YearRemodAdd 1994.0  MasVnrArea 0.0  BsmtFinSF1 383.5  BsmtFinSF2 0.0  BsmtUnfSF 477.5  TotalBsmtSF 991.5 | 1stFlrSF 1087.0  2ndFlrSF 0.0  LowQualFinSF 0.0  GrLivArea 1464.0  BsmtFullBath 0.0  BsmtHalfBath 0.0  FullBath 2.0  HalfBath 0.0  BedroomAbvGr 3.0  KitchenAbvGr 1.0  TotRmsAbvGrd 6.0  Fireplaces 1.0  GarageYrBlt 1980.0 | GarageCars 2.0  GarageArea 480.0  WoodDeckSF 0.0  OpenPorchSF 25.0  EnclosedPorch 0.0  3SsnPorch 0.0  ScreenPorch 0.0  PoolArea 0.0  MiscVal 0.0  MoSold 6.0  YrSold 2008.0  SalePrice 163000.0 |

#### b) Dữ liệu để test:

Giá trinh trung bình (mean):

|  |  |  |
| --- | --- | --- |
| MSSubClass 57.378341  LotFrontage 68.580357  LotArea 9819.161069  OverallQual 6.078821  OverallCond 5.553804  YearBuilt 1971.357779  YearRemodAdd 1983.662783  MasVnrArea 100.709141  BsmtFinSF1 439.203704  BsmtFinSF2 52.619342  BsmtUnfSF 554.294925  TotalBsmtSF 1046.117970 | BsmtHalfBath 0.065202  FullBath 1.570939  HalfBath 0.377656  BedroomAbvGr 2.854010  KitchenAbvGr 1.042495  TotRmsAbvGrd 6.385195  Fireplaces 0.581220  GarageYrBlt 1977.721217  GarageCars 1.766118  GarageArea 472.768861  WoodDeckSF 93.174777  OpenPorchSF 48.313914 | EnclosedPorch 24.243317  3SsnPorch 1.794380  ScreenPorch 17.064428  PoolArea 1.744345  MiscVal 58.167923  MoSold 6.104181  YrSold 2007.769705  LowQualFinSF 3.543523  GrLivArea 1486.045922  BsmtFullBath 0.434454  1stFlrSF 1156.534613  2ndFlrSF 325.967786 |

Giá trị độ lệch chuẩn (std):

|  |  |  |
| --- | --- | --- |
| MSSubClass 42.746880  LotFrontage 22.376841  LotArea 4955.517327  OverallQual 1.436812  OverallCond 1.113740  YearBuilt 30.390071  YearRemodAdd 21.130467  MasVnrArea 177.625900  BsmtFinSF1 455.268042  BsmtFinSF2 176.753926  BsmtUnfSF 437.260486  TotalBsmtSF 442.898624 | GrLivArea 485.566099  BsmtFullBath 0.530648  BsmtHalfBath 0.252468  FullBath 0.555190  HalfBath 0.503017  BedroomAbvGr 0.829788  KitchenAbvGr 0.208472  TotRmsAbvGrd 1.508895  Fireplaces 0.647420  GarageYrBlt 26.431175  GarageCars 0.775945  1stFlrSF 398.165820 | EnclosedPorch 67.227765  3SsnPorch 20.207842  ScreenPorch 56.609763  PoolArea 30.491646  MiscVal 630.806978  MoSold 2.722432  YrSold 1.301740  GarageArea 217.048611  WoodDeckSF 127.744882  OpenPorchSF 68.883364  LowQualFinSF 44.043251  2ndFlrSF 420.610226 |

Giá trị trung vị (median)

|  |  |  |
| --- | --- | --- |
| MSSubClass 50.0  LotFrontage 67.0  LotArea 9399.0  OverallQual 6.0  OverallCond 5.0  YearBuilt 1973.0  YearRemodAdd 1992.0  MasVnrArea 0.0  BsmtFinSF1 350.5  BsmtFinSF2 0.0  BsmtUnfSF 460.0  TotalBsmtSF 988.0 | 1stFlrSF 1079.0  2ndFlrSF 0.0  LowQualFinSF 0.0  GrLivArea 1432.0  BsmtFullBath 0.0  BsmtHalfBath 0.0  FullBath 2.0  HalfBath 0.0  BedroomAbvGr 3.0  KitchenAbvGr 1.0  TotRmsAbvGrd 6.0  Fireplaces 0.0 | GarageArea 480.0  WoodDeckSF 0.0  OpenPorchSF 28.0  EnclosedPorch 0.0  3SsnPorch 0.0  ScreenPorch 0.0  PoolArea 0.0  MiscVal 0.0  MoSold 6.0  YrSold 2008.0  GarageCars 2.0  GarageYrBlt 1979.0 |

### 1.3. Giá trị xuất hiện nhiều nhất (mode), ít nhất.

#### a) Bộ dữ liệu dùng để train

|  |  |
| --- | --- |
| MasVnrType less: BrkCmn most: None  MasVnrArea less: 442.0 most: 0.0  ExterQual less: Fa most: TA  ExterCond less: Po most: TA  Foundation less: Wood most: PConc  BsmtQual less: Fa most: TA  BsmtCond less: Po most: TA  BsmtExposure less: Mn most: No  BsmtFinType1 less: LwQ most: GLQ  BsmtFinSF1 less: 771.0 most: 0.0  BsmtFinType2 less: GLQ most: Unf  BsmtFinSF2 less: 750.0 most: 0.0  BsmtUnfSF less: 1211.0 most: 0.0  TotalBsmtSF less: 1075.0 most: 0.0  Heating less: Grav most: GasA  HeatingQC less: Po most: Ex  CentralAir less: N most: Y  Electrical less: FuseP most: SBrkr  1stFlrSF less: 2048 most: 864  2ndFlrSF less: 760 most: 0  LowQualFinSF less: 80 most: 0  GrLivArea less: 2046 most: 864  BsmtFullBath less: 3.0 most: 0.0  BsmtHalfBath less: 2.0 most: 0.0  FullBath less: 0 most: 2  HalfBath less: 2 most: 0  BedroomAbvGr less: 0 most: 3  KitchenAbvGr less: 0 most: 1  KitchenQual less: Fa most: TA  TotRmsAbvGrd less: 13 most: 6  Functional less: Sev most: Typ  Fireplaces less: 4 most: 0  FireplaceQu less: Ex most: Gd  GarageType less: CarPort most: Attchd  GarageYrBlt less: 1916.0 most: 2005.0  GarageFinish less: Fin most: Unf  GarageCars less: 5.0 most: 2.0  GarageArea less: 353.0 most: 0.0  GarageQual less: Po most: TA  GarageCond less: Ex most: TA | MSSubClass less: 150 most: 20  MSZoning less: RH most: RL  LotFrontage less: 131.0 most: 60.0  LotArea less: 12291 most: 9600  Street less: Grvl most: Pave  Alley less: Pave most: Grvl  LotShape less: IR3 most: Reg  LandContour less: Low most: Lvl  Utilities less: AllPub most: AllPub  LotConfig less: FR3 most: Inside  LandSlope less: Sev most: Gtl  Neighborhood less: Blueste most: NAmes  Condition1 less: RRNn most: Norm  Condition2 less: PosN most: Norm  BldgType less: 2fmCon most: 1Fam  HouseStyle less: 1.5Unf most: 1Story  OverallQual less: 1 most: 5  OverallCond less: 2 most: 5  YearBuilt less: 1879 most: 2005  YearRemodAdd less: 1982 most: 1950  RoofStyle less: Shed most: Gable  RoofMatl less: WdShngl most: CompShg  Exterior1st less: AsphShn most: VinylSd  Exterior2nd less: AsphShn most: VinylSd  PavedDrive less: P most: Y  WoodDeckSF less: 209 most: 0  OpenPorchSF less: 742 most: 0  EnclosedPorch less: 138 most: 0  3SsnPorch less: 86 most: 0  ScreenPorch less: 164 most: 0  PoolArea less: 144 most: 0  PoolQC less: Gd most: Ex  Fence less: MnWw most: MnPrv  MiscFeature less: Othr most: Shed  MiscVal less: 490 most: 0  MoSold less: 12 most: 6  YrSold less: 2010 most: 2007  SaleType less: Con most: WD  SaleCondition less: AdjLand most: Normal |

#### b) Bộ dữ liệu dùng để test

|  |  |
| --- | --- |
| MSSubClass less: 40 most: 20  MSZoning less: C (all) most: RL  LotFrontage less: 137.0 most: 60.0  LotArea less: 8190 most: 7200  Street less: Grvl most: Pave  Alley less: Pave most: Grvl  LotShape less: IR3 most: Reg  LandContour less: Low most: Lvl  Utilities less: NoSeWa most: AllPub  LotConfig less: FR3 most: Inside  LandSlope less: Sev most: Gtl  Neighborhood less: Blueste most: NAmes  Condition1 less: RRNe most: Norm  Condition2 less: RRAe most: Norm  BldgType less: 2fmCon most: 1Fam  HouseStyle less: 2.5Fin most: 1Story  OverallQual less: 1 most: 5  OverallCond less: 1 most: 5  YearBuilt less: 1872 most: 2006  YearRemodAdd less: 1951 most: 1950  RoofStyle less: Shed most: Gable  RoofMatl less: ClyTile most: CompShg  Exterior1st less: AsphShn most: VinylSd  Exterior2nd less: Other most: VinylSd  MasVnrType less: BrkCmn most: None  MasVnrArea less: 621.0 most: 0.0  ExterQual less: Fa most: TA  ExterCond less: Po most: TA  Foundation less: Wood most: PConc  BsmtQual less: Fa most: TA  BsmtCond less: Po most: TA  BsmtExposure less: Mn most: No  BsmtFinType1 less: LwQ most: Unf  BsmtFinSF1 less: 1904 most: 0  BsmtFinType2 less: GLQ most: Unf  BsmtFinSF2 less: 1474 most: 0  BsmtUnfSF less: 2046 most: 0  SaleType less: Con most: WD  SaleCondition less: AdjLand most: Normal  SalePrice less: 34900 most: 140000 | TotalBsmtSF less: 2046 most: 0  Heating less: Floor most: GasA  HeatingQC less: Po most: Ex  CentralAir less: N most: Y  Electrical less: Mix most: SBrkr  1stFlrSF less: 2053 most: 864  2ndFlrSF less: 1872 most: 0  LowQualFinSF less: 384 most: 0  GrLivArea less: 2054 most: 864  BsmtFullBath less: 3 most: 0  BsmtHalfBath less: 2 most: 0  FullBath less: 0 most: 2  HalfBath less: 2 most: 0  BedroomAbvGr less: 8 most: 3  KitchenAbvGr less: 0 most: 1  KitchenQual less: Fa most: TA  TotRmsAbvGrd less: 2 most: 6  Functional less: Sev most: Typ  Fireplaces less: 3 most: 0  FireplaceQu less: Po most: Gd  GarageType less: 2Types most: Attchd  GarageYrBlt less: 1906.0 most: 2005.0  GarageFinish less: Fin most: Unf  GarageCars less: 4 most: 2  GarageArea less: 1418 most: 0  GarageQual less: Ex most: TA  GarageCond less: Ex most: TA  PavedDrive less: P most: Y  WoodDeckSF less: 215 most: 0  OpenPorchSF less: 119 most: 0  EnclosedPorch less: 552 most: 0  3SsnPorch less: 508 most: 0  ScreenPorch less: 480 most: 0  PoolArea less: 480 most: 0  PoolQC less: Ex most: Gd  Fence less: MnWw most: MnPrv  MiscFeature less: TenC most: Shed  MiscVal less: 54 most: 0  MoSold less: 2 most: 6  YrSold less: 2010 most: 2009 |

### 1.4. Tỉ lệ các giá trị bị thiếu của các thuộc tính

#### a) Bộ dữ liệu dùng để train:

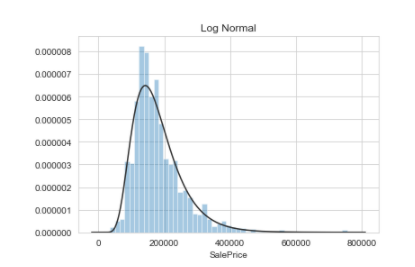
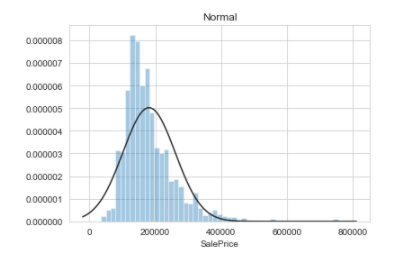
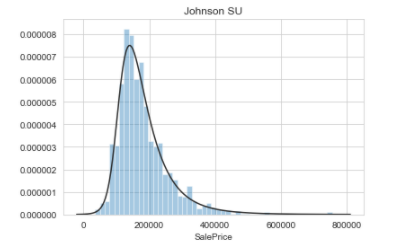
|  |  |
| --- | --- |
| PoolQC 99.52  MiscFeature 96.30  Alley 93.77  Fence 80.75  FireplaceQu 47.26  LotFrontage 17.74  GarageCond 5.55  GarageType 5.55  GarageYrBlt 5.55  GarageFinish 5.55 | GarageQual 5.55  BsmtExposure 2.60  BsmtFinType2 2.60  BsmtFinType1 2.53  BsmtCond 2.53  BsmtQual 2.53  MasVnrArea 0.55  MasVnrType 0.55  Electrical 0.07  ….. = 0.00 |

#### b) Bộ dữ liệu dùng để test:

|  |  |
| --- | --- |
| PoolQC 99.79  MiscFeature 96.50  Alley 92.67  Fence 80.12  FireplaceQu 50.03  LotFrontage 15.56  GarageCond 5.35  GarageQual 5.35  GarageYrBlt 5.35  GarageFinish 5.35  GarageType 5.21  BsmtCond 3.08  BsmtQual 3.02  BsmtExposure 3.02  BsmtFinType1 2.88  BsmtFinType2 2.88  MasVnrType 1.10 | MasVnrArea 1.03  MSZoning 0.27  BsmtHalfBath 0.14  Utilities 0.14  Functional 0.14  BsmtFullBath 0.14  BsmtFinSF2 0.07  BsmtFinSF1 0.07  Exterior2nd 0.07  BsmtUnfSF 0.07  TotalBsmtSF 0.07  SaleType 0.07  Exterior1st 0.07  KitchenQual 0.07  GarageArea 0.07  GarageCars 0.07  ….. = 0.00 |

## 2. Trực quan hóa dữ liệu

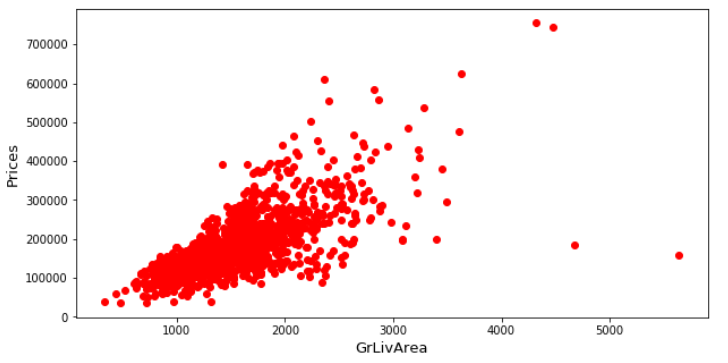
### 2.1 Biểu diễn giá trị SalePrice:

****

Các giá trị SalePrice không tuân theo phân phối chuẩn, vì vậy trước khi thực hiện hội quy, nó phải được chuyển đổi. Trong khi chuyển đổi Log Normal hoạt động khá tốt, phù hợp nhất là phân phối Johnson không bị ràng buộc

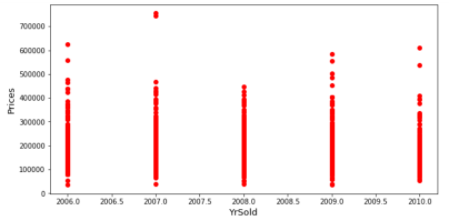
### 2.2 Biểu diễn theo các thuộc tính đối với giá nhà

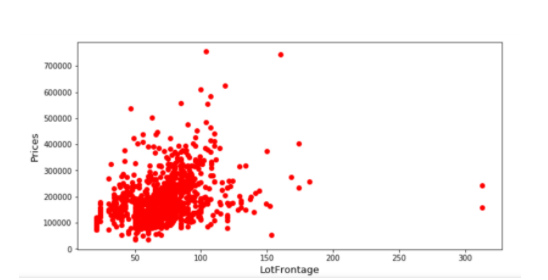
Từ đồ thị của một số thuộc tính, có thể thấy mối tương quan như sau:

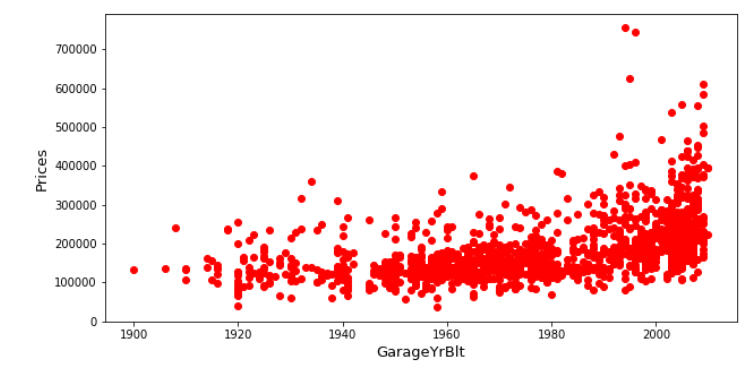


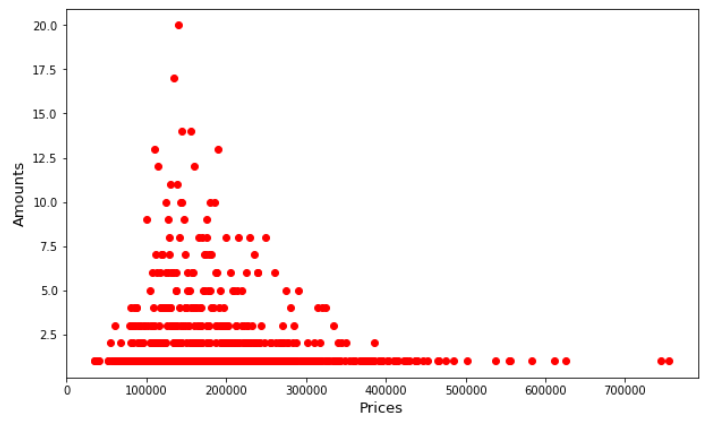
Từ đồ thị có thể thấy thuộc tính GrLivArea càng cao thì giá càng lớn, một số thuộc tính có đồ thị tương tự là: 1stFlrSF, 2ndFlrSF( đối với các giá trị !=0), Garage Area( phân bố loạn hơn). Từ những thuộc tính như thế này, khi chạy model xong có thể kiểm tra trọng số thuộc tính này so với các thuộc tính khác.

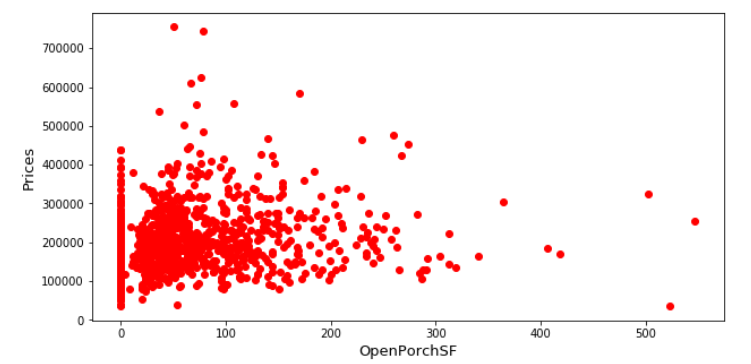
Đối với một số thuộc tính khác, đồ thị thể hiện sự rời rạc như YrSold, hoặc phân bố chưa thể hiện đc nhiều gì, đồ thị sẽ trông như sau:

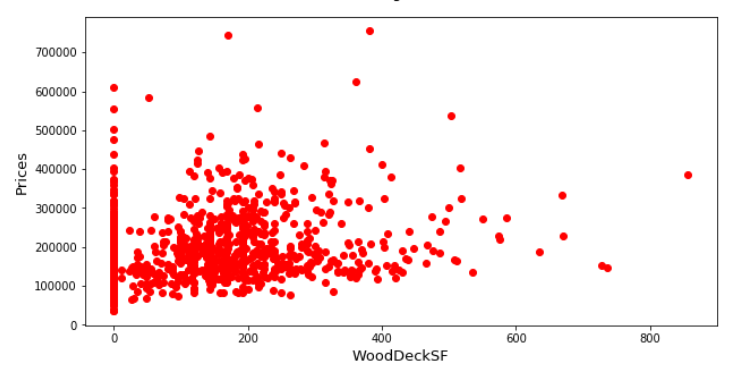


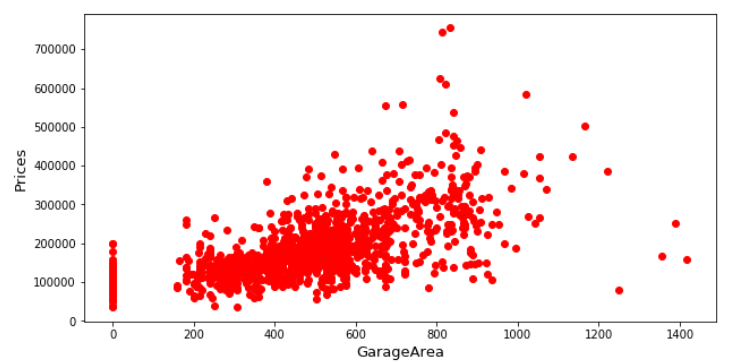












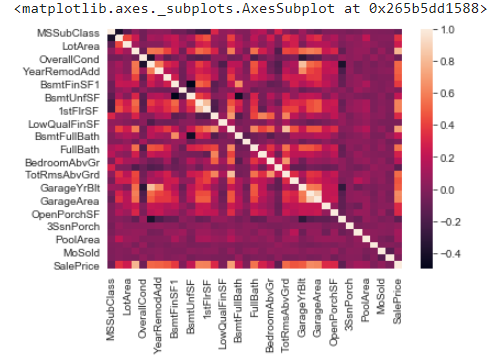
Các thuộc tính so sánh còn lại có trong code Visualize

### 2.3 Biểu diễn độ tương quan của các thuộc tính dữ liệu

Từ bộ dữ liệu ta chia được dữ liệu thành 2 phần theo :

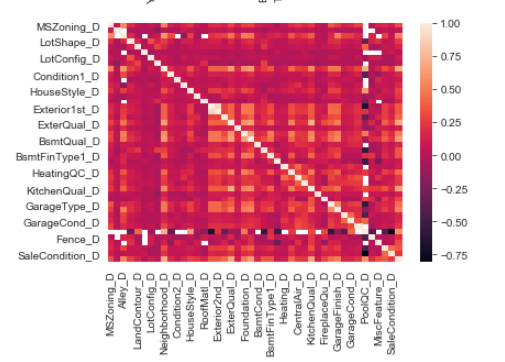
* Định lượng (int, float và gồm 38 thuộc tính):

Qua các thuộc tính biểu diễn dữ liệu đánh giá phản ánh các thuộc tính dữ liệu với nhau qua bảng này

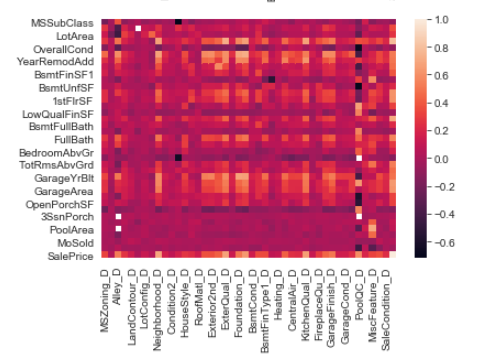


* Định tính (object và gồm 43 thuộc tính):

Qua các thuộc tính dưới dạng object này em biểu diễn với mỗi giá trị trong object thành 1 số biểu diễn sau đó tạo thành ma trận đánh giá các thuộc tính dữ liệu với nhau qua bảng sau:

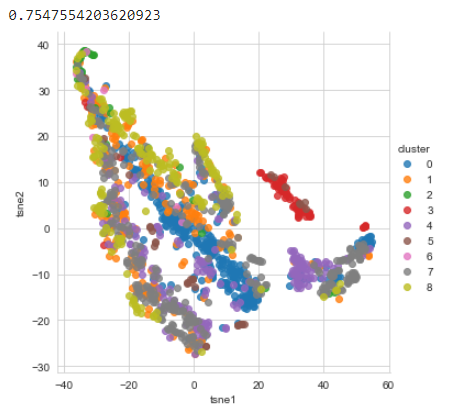


Sau đó em biểu diễn đánh giá độ tương phản giữa phần thuộc tính và phần định lượng với nhau:



### 2.4 Biểu diễn các điểm trên đồ thị hai chiều

Biểu diễn các điểm dữ liệu với dưới dạng hai chiều và phần thành các 9 cụm (thuật toán Kmeans):

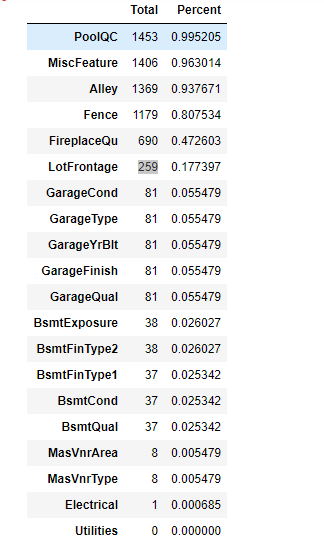


## 3. Xử lý dữ liệu bị thiếu

Dựa vào bảng dữ liệu đã cho, ta có thể thấy có khá nhiều thuộc tính bị thiếu dữ liệu. Dưới đây là bảng số lượng và tỉ lệ các dữ liệu bị thiếu của các thuộc tính:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Thuộc tính | Số lượng | Tỉ lệ (%) | Thuộc tính | Số lượng | Tỉ lệ (%) |
| PoolQC | 2909 | 0.996574 | **BsmtFinType2** | 80 | 0.027407 |
| MiscFeature | 2814 | 0.964029 | **BsmtFinType1** | 79 | 0.027064 |
| Alley | 2721 | 0.932169 | **MasVnrType** | 24 | 0.008222 |
| Fence | 2348 | 0.804385 | **MasVnrArea** | 23 | 0.007879 |
| SalePrice | 1459 | 0.499829 | **MSZoning** | 4 | 0.001370 |
| FireplaceQu | 1420 | 0.486468 | **BsmtFullBath** | 2 | 0.000685 |
| LotFrontage | 486 | 0.166495 | **BsmtHalfBath** | 2 | 0.000685 |
| GarageFinish | 159 | 0.054471 | **Utilities** | 2 | 0.000685 |
| GarageCond | 159 | 0.054471 | **Functional** | 2 | 0.000685 |
| GarageQual | 159 | 0.054471 | **Electrical** | 1 | 0.000343 |
| GarageYrBlt | 159 | 0.054471 | **Exterior2nd** | 1 | 0.000343 |
| GarageType | 157 | 0.053786 | **KitchenQual** | 1 | 0.000343 |
| BsmtCond | 82 | 0.028092 | **Exterior1st** | 1 | 0.000343 |
| BsmtExposure | 82 | 0.028092 | **GarageCars** | 1 | 0.000343 |
| BsmtQual | 81 | 0.027749 | **TotalBsmtSF** | 1 | 0.000343 |

Xem thông chi tiết từng file như sau:



### 3.1 Loại bỏ một số trường dữ liệu

Ta thấy số lượng cũng như tỉ lệ thiếu của thuộc tính PoolQC là rất lớn. Điều này là dễ hiểu bởi vì số lượng các căn nhà có bể bơi là không nhiều. Tương tự các thuộc tính Alley; Loại hẻm đi vào , Fence: Chất lượng rào, MiscFeature: Tính năng khác(thang máy, teniss,..)

Vì do lượng dữ liệu mất lớn ( > 1000) nên bọn em drop các trường bị mất nhiều dữ liệu này

### 3.2 Lấy giá trị trung bình ở một số trường dữ liệu

Ta thấy các trường dữ liệu sau :

+ LotFrontage: cách đường bn

+ BsmtFinSF1: xếp hạng khu vực hoàn thành tầng hầm số m2 hoàn thành

+ BsmtFinSF2: xếp hạng khu vực hoàn thành tầng hầm nếu có 2

+ BsmtUnfSF: feet vuông chưa hoàn thành

+TotalBsmtSF: tổng số feet vuông của hầm

+GarageCars: Chưa được mấy xe

+GarageArea: kích thước gara

🡺 Đối nhiều dữ liệu bằng thực nghiệm bọn em quyết định lấy ra trị trung bình cho các giá trị bị thiếu đối với trường này

### 3.3 Lấy giá trị tần suất xuất hiện nhiều nhất với các trường

Ta thấy các trường dữ liệu sau:

+ BsmtCond: tình trạng chung tầng hầm

+ BsmtQual: đánh giá chiều cao tầng hầm

+ FireplaceQu: chất lượng lò sưởi

+ GarageType: vị trí gara

+ GarageYrBlt: năm gara xây

+ GarageFinish: hoàn thành nội thất

+ GarageQual: chất lượng

+ GarageCond: điều kiện nhà để xe

+ MasVnrType: gạch hay kết dính

+ MasVnrArea: diện tích cái thằng trên tính bằng số feet vuông

+ BsmtExposure: tường đi bộ hoặc tường vườn

+ BsmtFinType2: xếp hàng khu vực hoàn thành tầng hầm

+ Utilities: tiện ích có sẵn

+ Exterior1st: Ngoại thất( bê tông, xi măng, kim loại,..)

+ Exterior2nd: vẫn là ngoại thất

+ BsmtFinType1: xếp hạng khu vực hoàn thành tầng hầm

+ BsmtFullBath: tầng hầm đẩy đủ phòng tắm

+ BsmtHalfBath: tầng hầm 1 nửa phòng tắm

+ KitchenQual: đánh giá

+ Functional: chức năng gia đình

+ GarageCars: chứa được mấy xe

+ GarageArea: kích thước gara

+ SaleType: kiểu bán

🡺 Ta chọn giá trị có số lần xuất hiện nhiều nhất để gán cho các giá trị còn thiếu.

## 4. Chuẩn hóa dữ liệu

### 4.1 Chia dữ liệu làm 2 phần

+ dữ liệu thuộc kiểu object : data\_cat

+ dữ liệu thuộc kiểu int64,float64: data\_num

### 4.2 Đối với phần dữ liệu thuộc kiểu data\_num

Giữ nguyên phần dữ kiểu dữ liệu số sau khi đã xử lý xong phần dữ liệu bị missing

### 4.3 Đối với phần dữ liệu thuộc kiểu data\_cat

Sử dụng xây dựng hàm category\_onehot\_multcols để chuẩn hóa kiểu dữ liệu object này bằng việc sữ dụng thư viện pandas.get\_dummies()

Hàm pandas.get\_dummies() : Chuyển đổi biến phân loại thành biến giả/ chỉ số

### 4.4 Kết hợp tạo ra tập train và test

Loại bỏ các trường bị trùng lặp sau khi chuẩn hóa tập dữ liệu data\_cat Phân chia:

+ Tập dữ liệu train có kích thước và số chiều lần lượt là: 1422 và 175

+Tập dữ liệu test có kích thước và số chiều lần lượt là:1459 và 175

# IV. Thực nghiệm

## Thực hiện

Dữ liệu được chia thành hai phần, một phần để huấn luyện, một phần để kiểm tra theo tỉ lệ 6:4.

Nhóm tìm hiểu một số thuật toán và thực hiện cài đặt để kiểm thử và đánh giá hiệu quả.

Các thuật toán nhóm sử dụng : Linear Regression, LASSO Regression, Ridge Regression, Random Forest và ANN.

Với các giải thuật LASSO Regression và Ridge Regression, hằng số chính quy hóa là 0.5

Với giải thuật Random Forest thì nhóm chọn n\_estimator = 100 và random\_state = 0

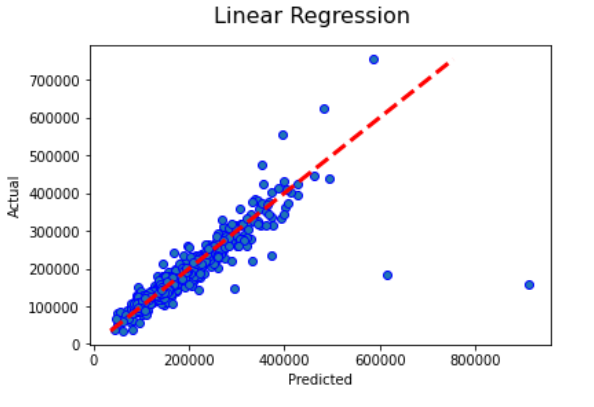
Ngoài ra, chúng em còn sử dụng các thuật toán trên để dự đoán kết quả cho một bộ dữ liệu của Kaggle, và nộp kết quả lên để chấm điểm so sánh.

Kaggle sử dụng RMSLE(Root Mean Squared Log Error) để đánh giá.

## Kết quả

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | accuracy  (test)(%) | mae | mse | r2 | Kaggle score |
| Linear Regression | 70.61 | 20693.61080378823 | 1821066702.928624 | 0.7061075556367518 | 0.53653 |
| LASSO Regression | 68.05 | 20838.161724121885 | 2145976734.0605655 | 0.6493121011191259 | 0.54204 |
| Ridge Regression | 81.41 | 20593.82511259496 | 1133951680.6996555 | 0.8140856060921247 | 0.22520 |
| Random Forest | 92.08 | 11576.071739894553 | 510448023.62799275 | 0.9208421607275935 | 0.15840 |
| ANN | 87.14 | 24858.004427998683 | 1891354845.4870186 | 0.7443289735015681 | 0.18562 |

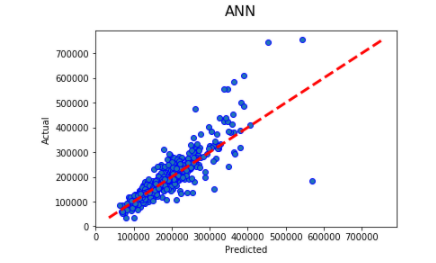
Tương quan giữa kết quả dự đoán được và giá trị thực tế của các thuật toán:



# 

# 

# 



# 

# 

# V. Nhận xét đánh giá

Thuật toán Random Forest đạt hiệu quả tốt nhất. Điều này là dễ hiểu vì Random Forest là một trong những thuật toán có hiệu quả tốt trên nhiều bài toán.

Thuật toán LASSO Regression khá tồi. Điều này có thể bởi việc áp dụng L1 có thể khiến một số đặc trưng quan trọng bị triệt tiêu.

Các thuật toán sử dụng còn khá đơn giản,cùng với việc xử lý dữ liệu còn phụ thuộc vào kinh nghiệm nên hiệu quả có thể chưa tốt.

# VI. Quá trình làm việc

Phân chia công việc

|  |  |
| --- | --- |
| Công việc | Phân công |
| Thống kê dữ liệu trên tệp dữ liệu House Prices của Kaggle, đưa ra phương án xử lý dữ liệu bị thiếu | Cường |
| Xử lý dữ liệu bị thiếu, chuẩn hóa dữ liệu | Đạt |
| Triển khai các thuật toán Linear Regression, Lasso Regression, Ridge Regression | Phúc |
| Triển khai các thuật toán Random Forest | Thúy |
| Triển khai các thuật toán Artificial Neural Network | Đạt |
| Tổng hợp kết quả, so sánh và báo cáo | Cả nhóm |

# VII. Tài Liệu tham khảo

1. Slide Nhập môn Học máy và Khai phá dữ liệu, Thân Quang Khoát, ĐH Bách khoa Hà Nội
2. Slide Nhập môn Học máy và khai phá dữ liệu, Nguyễn Nhật Quang, ĐH Bách khoa Hà Nội
3. Machine Learning cơ bản, Vũ Hữu Tiệp, NXB Lao Động
4. Bộ dữ liệu từ cuộc thi House Prices - Advanced Regression Techniques của Kaggle: <https://www.kaggle.com/c/house-prices-advanced-regression-techniques/>
5. Sci-kit-learn documentation