

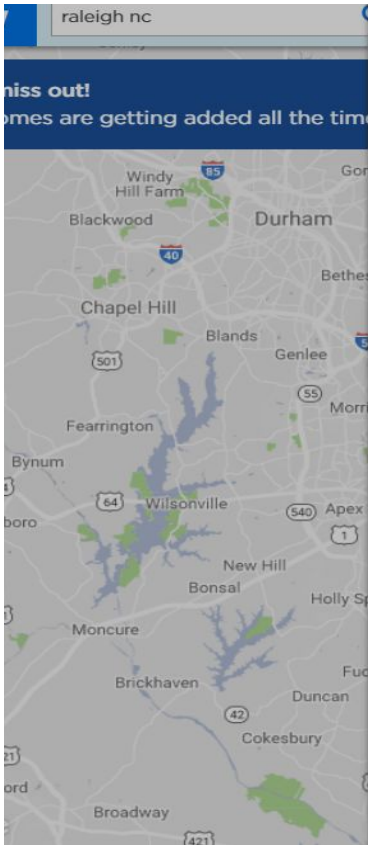
# House Price Prediction in Iowa State

By,

Akshit Meghawat, ammeghaw

Harika Malapaka, hsmalapa

Sanya Kathuria, skathur2



## INTERIOR FEATURES

### Bedrooms

Beds: 7

### Bathrooms

Baths: 8 full, 4 half

### Flooring

Floor size: 18,398 sqft

Flooring: Carpet, Hardwood, Tile

### Other Interior Features

Fireplace

[View Virtual Tour](#)

## SPACES AND AMENITIES

### Spaces

Wet Bar

Pool

Fitness Center

Tennis Court

### Amenities

Elevator

Security System

## CONSTRUCTION

### Type and Style

Single Family

### Dates

Built in 1999

### Materials

Exterior material: Stone

## EXTERIOR FEATURES

### Patio

Porch

### Lot

Lot: 15.27 acres

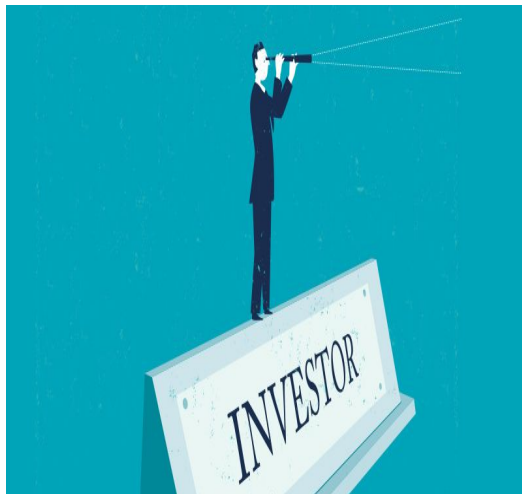
**Given Features  
of a House.**

**Predict it's  
Price?**



# Introduction and Motivation

## Business Point of View



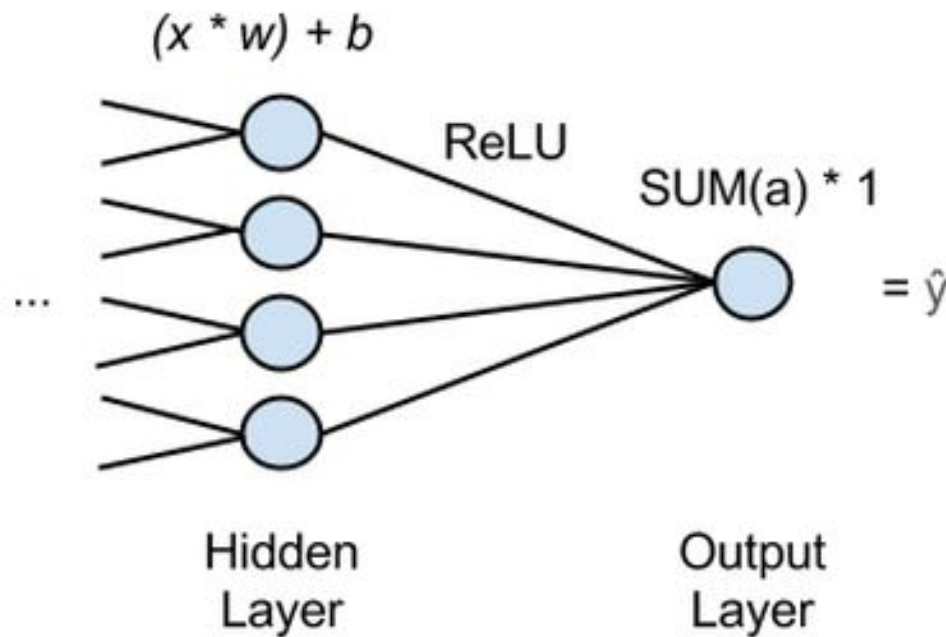
**How does he predict which house is the best?**

**Solution : Use this model**

# Class Point of View

Using Neural Networks With Regression

deeplearning4j



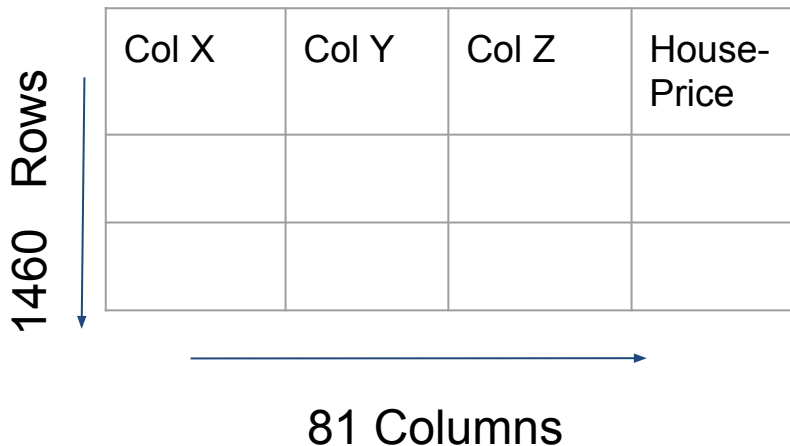
**Advanced  
Neural Nets**



# Dataset

- Target attribute :  
Sale Price of the house  
Ex : \$39050
- Dataset contains 80 attributes  
EX : Basement Quality : good

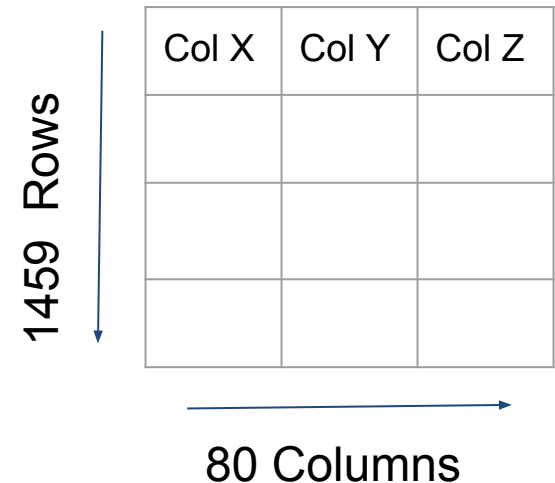
## Train Dataset



A diagram showing the structure of the Train Dataset. It consists of a table with 4 columns and 1460 rows. The columns are labeled 'Col X', 'Col Y', 'Col Z', and 'House-Price'. The first three columns are empty, and the fourth column contains the text 'House-Price'. A vertical arrow on the left indicates 1460 Rows, and a horizontal arrow at the bottom indicates 81 Columns.

Col X	Col Y	Col Z	House-Price

## Test Dataset

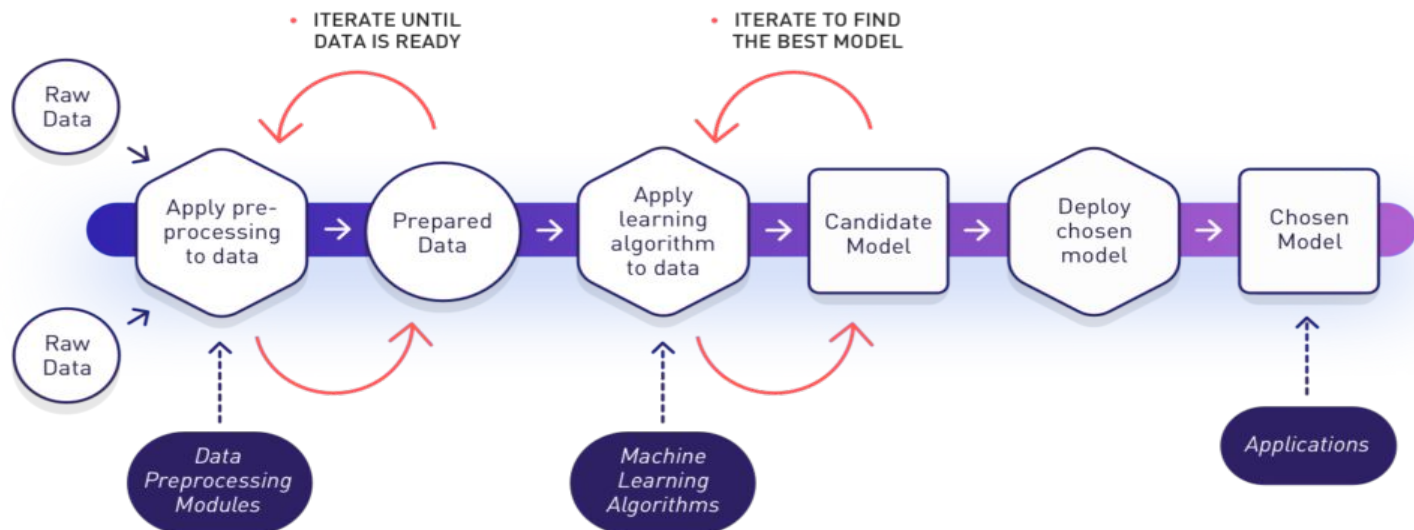


A diagram showing the structure of the Test Dataset. It consists of a table with 3 columns and 1459 rows. The columns are labeled 'Col X', 'Col Y', and 'Col Z'. All three columns are empty. A vertical arrow on the left indicates 1459 Rows, and a horizontal arrow at the bottom indicates 80 Columns.

Col X	Col Y	Col Z

# Literature Survey

- Advanced Machine learning
  - Data Cleaning
  - Data Preprocessing
  - Supervised Learning (Neural Network)
  - Regression



# Data Cleaning and Preprocessing

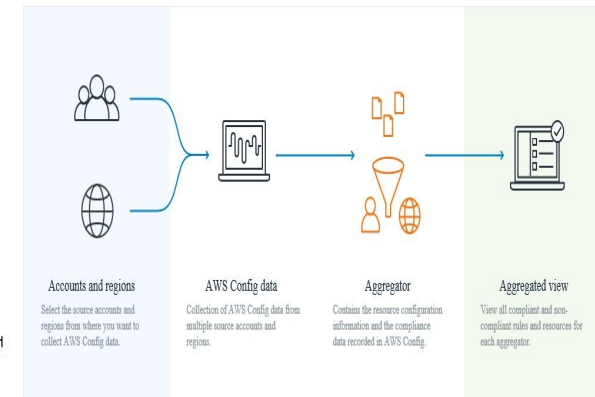
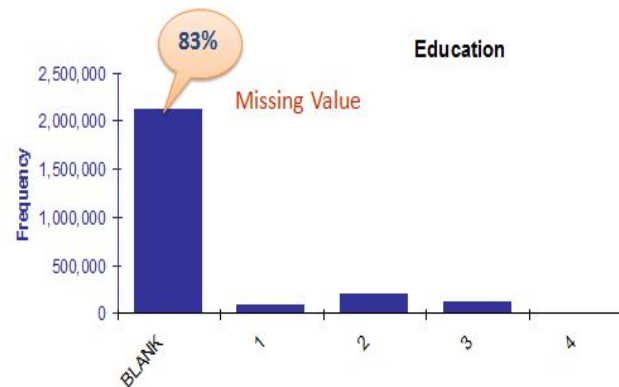
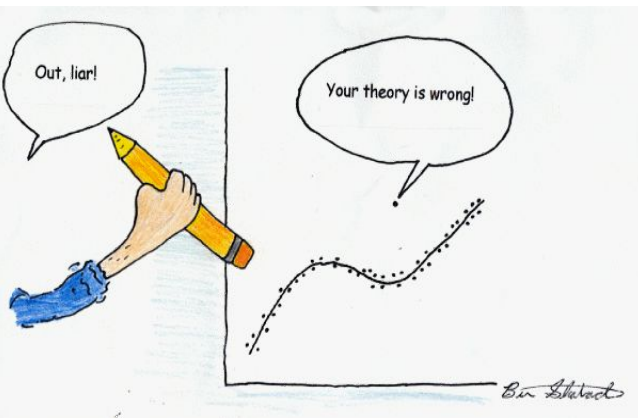
The below steps were applied on both train and test data :

Data Cleaning steps :

1. Removing Outliers

2. Handle Missing Values

3. Aggregations



**Newly Processes Data:**

**Train-set : 1417 rows, 67 columns**

**Test set : 1459 rows, 66 columns**

# Data Preprocessing

Data Preprocessing steps : Normalization (Min Max Scaler)

- Used only continuous features
- Used continuous and categorical features

**Min-Max Normalization**

Formula: 
$$\text{new\_value} = \frac{\text{value} - \text{min}}{\text{max} - \text{min}} \times (\text{new\_max} - \text{new\_min}) + \text{new\_min}$$

here  
 $\text{new\_min} = 0$   
 $\text{new\_max} = 1$   
 $\text{min\_A} = 13$   
 $\text{max\_A} = 27$

min-max for 13: 
$$\frac{13 - 13}{27 - 13} \times (1 - 0) + 0 = \frac{0}{14} \times 1 = 0$$

min-max for 17: 
$$\frac{17 - 13}{27 - 13} \times (1 - 0) + 0 = \frac{4}{14} \times 1 = 0.2857$$

min-max for 19: 
$$\frac{19 - 13}{27 - 13} \times (1 - 0) + 0 = \frac{6}{14} \times 1 = 0.4285$$

min-max for 27: 
$$\frac{27 - 13}{27 - 13} \times (1 - 0) + 0 = \frac{14}{14} \times 1 = 1$$

values	values after min-max normalization
13	0
17	0.2857
19	0.4285
27	1

**NORMALIZATION**

train-numerical.csv  
 train-categorical.csv  
 train-all.csv

test-numeric.csv  
 test-categorical.csv  
 test-all.csv




# Method: Linear Regression

$$Y = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n$$

Independent Variables (Features)

Dependent Variable (Target)



TotalBsmtSF	LowQualFinSF	...	Fireplaces	GarageYrBlt	GarageCars	MiscVal	MoSold	YrSold	SalePrice
0.249532	0.0	...	0.000000	0.978109	0.50	0.0	0.545455	0.50	0.239442
0.224891	0.0	...	0.000000	0.970149	0.25	0.0	0.727273	0.75	0.081664
0.392389	0.0	...	0.333333	0.998507	0.75	0.0	0.454545	0.25	0.238575
0.192140	0.0	...	0.000000	0.955721	0.25	0.0	0.454545	0.00	0.093800
0.283843	0.0	...	0.333333	0.990050	0.50	0.0	0.363636	0.25	0.251406

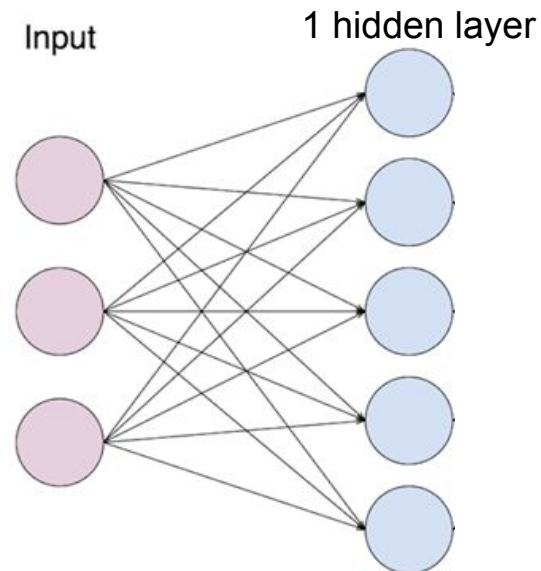
# Method: Neural Net Regression

Broadly speaking we experimented with multiple 'sizes' of networks.

- Deep network: 100, 50, 25, 12, 6
- Medium sized network: 50, 25, 12, 6, 3
- Small sized network: 20, 10, 5

**Shallow Net :**  
**Only 1 hidden layer**

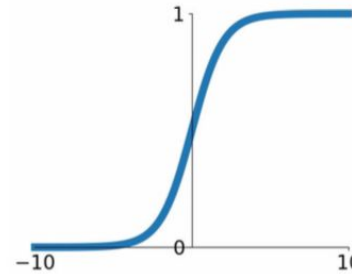
**Number of neurons : 5, 12, ....., 400**



# Activation Functions

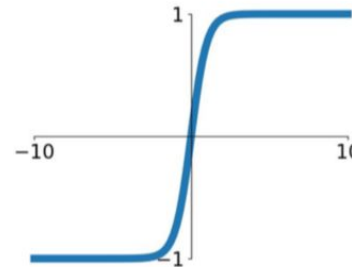
## Sigmoid

$$\sigma(x) = \frac{1}{1+e^{-x}}$$



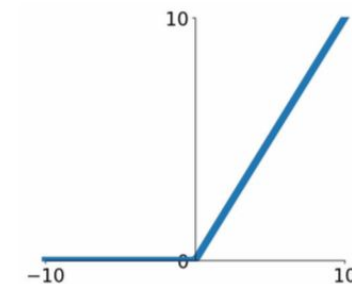
## tanh

$$\tanh(x)$$



## ReLU

$$\max(0, x)$$



# Results (1)

## Continuous Features:

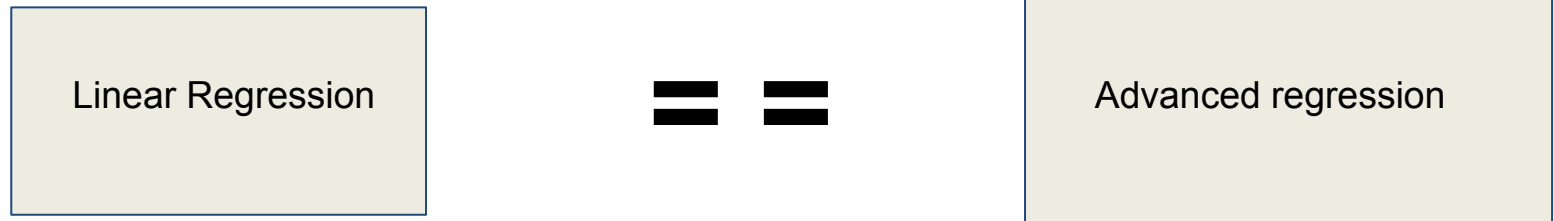
Regressor	MSE	Neural Net Layers	Number of Neurons in each layer
DNN Regression (complex)	0.0026	5	100, 50, 25, 12, 6
DNN Regression (medium)	0.0040	5	50, 25, 12, 6, 3
DNN Regression (small)	0.0038	3	20, 10, 5
Shallow Regression	0.0063	1	5
Shallow Regression	0.0035	1	12
Shallow Regression	0.0028	1	25
Shallow Regression	0.0038	1	50
Shallow Regression	0.0027	1	100
Shallow Regression	0.0026	1	200
Shallow Regression	0.0026	1	400
Linear regression	0.0028	—	—

## Results (2)

### Continuous and Categorical Features:

Regressor	MSE	Neural Net Layers	Number of Neurons in each layer
DNN Regression (complex)	0.0020	5	100, 50, 25, 12, 6
DNN Regression (medium)	0.0027	5	50, 25, 12, 6, 3
DNN Regression (small)	0.0016	3	20, 10, 5
Shallow Regression	0.0021	1	5
Shallow Regression	0.0018	1	12
Shallow Regression	0.0019	1	25
Shallow Regression	0.0019	1	50
Shallow Regression	0.0019	1	100
Shallow Regression	0.0019	1	200
Shallow Regression	0.0020	1	400
Linear regression	0.0024	—	—

# Comparison Of Advanced Regression with Simple Linear Regression



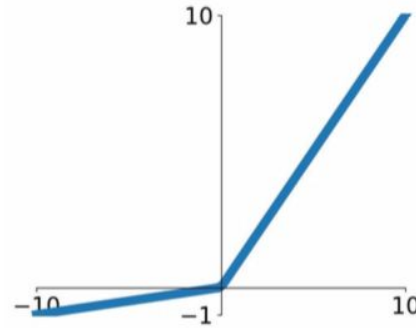
**Is their performance Comparable?**

**Answer : YES** (After using feature scaling)

**Then why use Advanced techniques?**

## Leaky ReLU

$$\max(0.1x, x)$$

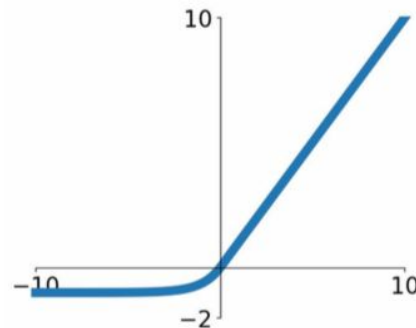


## Maxout

$$\max(w_1^T x + b_1, w_2^T x + b_2)$$

## ELU

$$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$$



# Conclusions

- 60 % of the work was to clean and preprocess the data.
- As we make the model complicated with more number of layers and neurons, the model might be over-fitted.
- Educated guesses would be good in selecting the number of layers and number of neurons
- Linear regression might not be the best approach.
- MSE value is lower when we use all features instead of only continuous values as features.



# Thank you!

**ANY  
QUESTIONS?**

