NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

SC1015 : Review Lecture

ML basics, Gradient descent and **Regression**

Dr Smitha K G

SC1015

Admin Announcements

1. Detailed solutions to the Lab Exercises will be posted every week after the Lab Week is over.
2. Mini-Project details will be posted by Week 6.

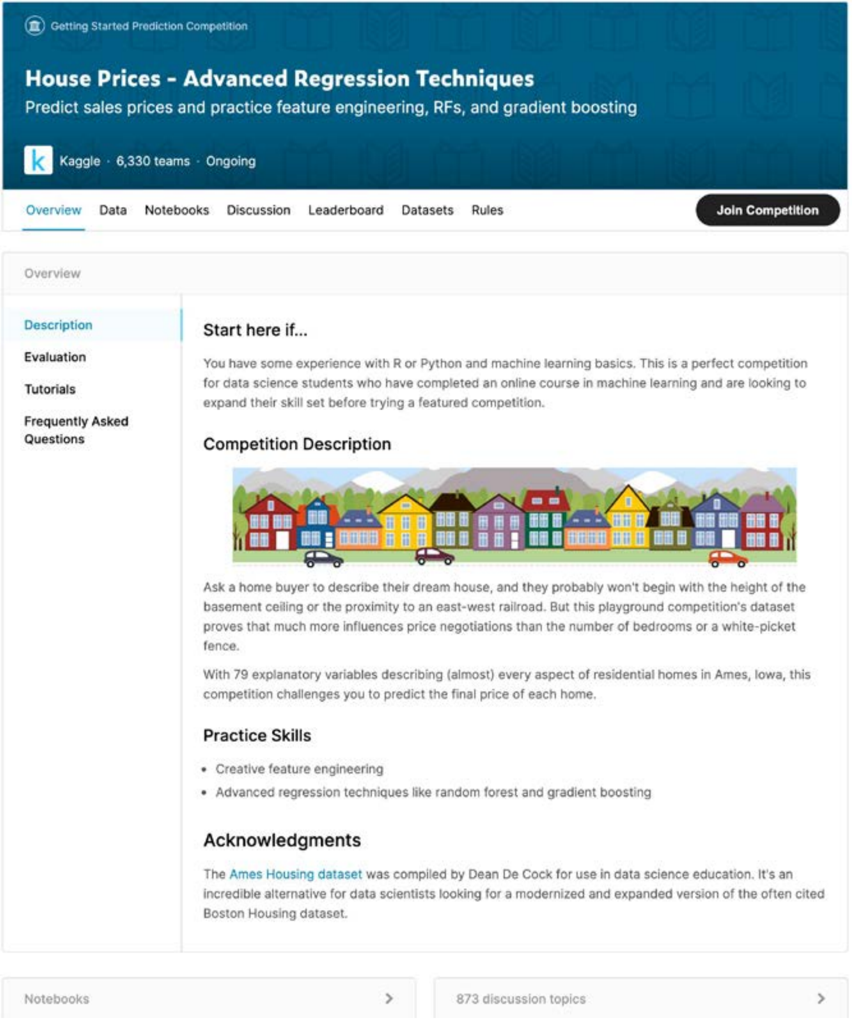
LAMS Completion Status

Module 1 Part 1 : Above 900 – Quiz solutions posted
Module 1 Part 2 : Above 875 – Quiz solutions posted
Module 2 Part 1 : Above 800 – Quiz solutions posted
Module 2 Part 2 : Above 650 – Complete by this week
Module 3 : Above 350 – Complete by Exercise 4 (W6)

Graded Lab Exercises in Weeks 4, 6, 7 – 5% each.
DS Theory Quiz in Recess Week : 8 March, Friday.

Two time slots 12.30 to 2pm and 2.30 to 4pm

Time slot for a lab group is fixed and we will let you know by end of this week



Getting Started Prediction Competition

House Prices - Advanced Regression Techniques

Predict sales prices and practice feature engineering, RFs, and gradient boosting

Kaggle · 6,330 teams · Ongoing

Overview Data Notebooks Discussion Leaderboard Datasets Rules Join Competition

Overview

Description
Evaluation
Tutorials
Frequently Asked Questions

Start here if...

You have some experience with R or Python and machine learning basics. This is a perfect competition for data science students who have completed an online course in machine learning and are looking to expand their skill set before trying a featured competition.

Competition Description

Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this playground competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.

With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition challenges you to predict the final price of each home.

Practice Skills

- Creative feature engineering
- Advanced regression techniques like random forest and gradient boosting

Acknowledgments

The Ames Housing dataset was compiled by Dean De Cock for use in data science education. It's an incredible alternative for data scientists looking for a modernized and expanded version of the often cited Boston Housing dataset.

Notebooks

873 discussion topics

Data Science

Machine Learning

Prediction : Numeric

Regression

Model : SalePrice = *f* (Variables)

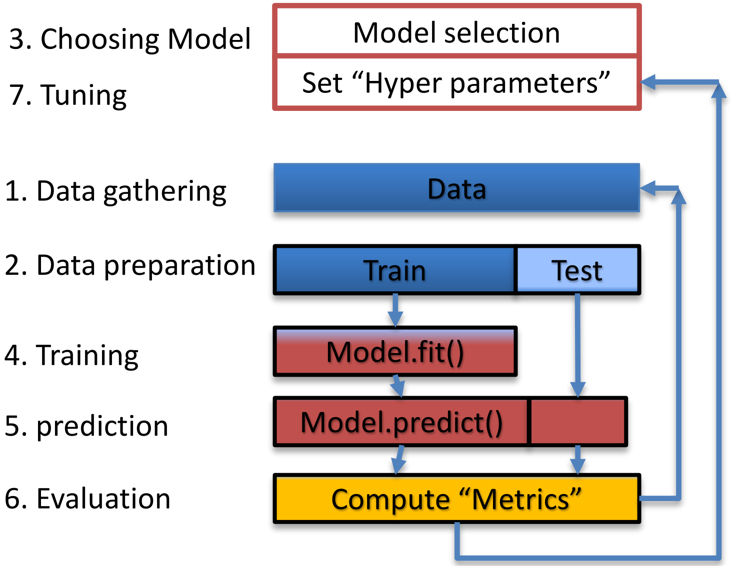
Given	Some Houses as Train Data
Learn	The Formula for SalePrice
Predict	Estimate SalePrice for Test

https://www.kaggle.com/c/house-prices-advanced-regression-techniques

Data Science

Machine learning

Basic Steps in Machine learning

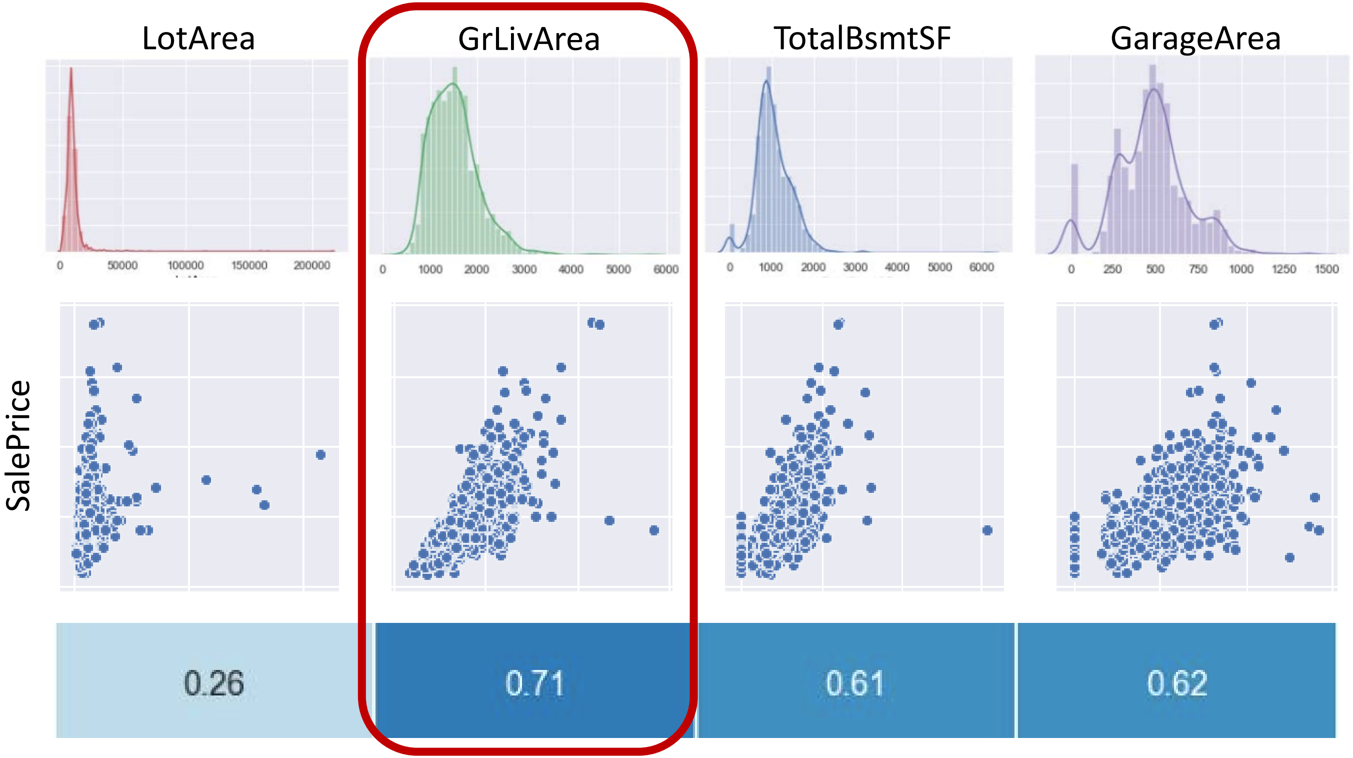


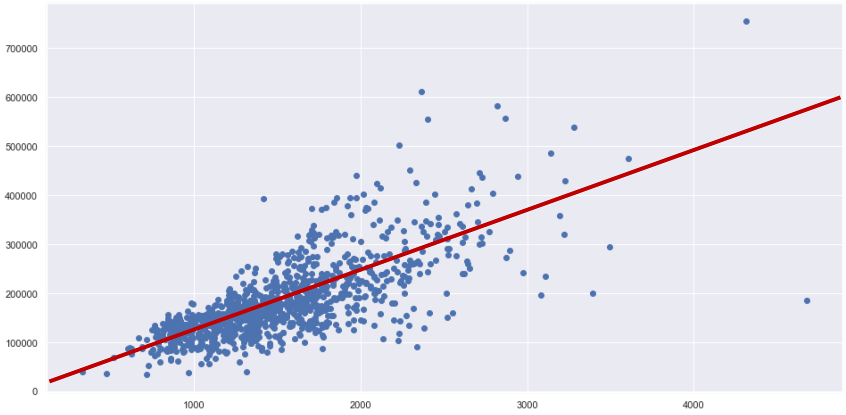
Building and evaluating a model

1. Prepare the ‘Train’ and ‘Test’ sets
2. Choose a “Model family” based on the problem
3. ‘Fit’ the Model on the ‘Train’ set
4. Predict the value of ‘Test’ using Trained model
5. Evaluate Performance Metrics for Model

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html
https://scikit-learn.org/stable/modules/classes.html#module-sklearn.linear_model
https://scikit-learn.org/stable/modules/classes.html#sklearn-metrics-metrics

The 7 steps of Machine Learning : https://youtu.be/nKW8Ndu7Mjw


LotArea
GrLivArea
TotalBsmtSF
GarageArea
0
50000
100000
150000
200000
0
1000
2000
3000
4000
5000
6000
0
1000
2000
3000
4000
5000
6000
0
250
500
750
1000
1250
1500
SalePrice
0.26
0.71
0.61
0.62



Data Science

Uni-Variate Regression

Model (linear) : $\boldsymbol{y} = \boldsymbol{a}\boldsymbol{x} + \boldsymbol{b} + \epsilon$

Fit on the Train Set and evaluated on Test Set

Training = Minimization of Cost Function

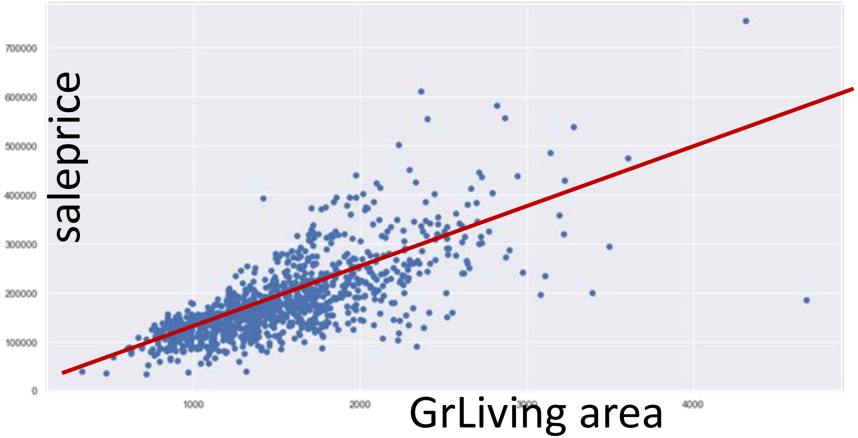
$$J(a,b) = \sum (y - \hat{y})^2 = \sum (y - ax - b)^2$$

Steps automatically done by model.fit()

Guess parameters ***a*** and ***b*** in the model
Predict the values of ***y*** in the Train Data
Calculate the ***errors*** compared to actual
Tune parameters (***a***, ***b***) to minimize ***Cost***

Pause and Ponder : Why this specific form of "Cost Function" for Linear Regression?

Uni-Variate Regression



Steps automatically done by model.fit()

Guess parameters ***a*** and ***b*** in the model Predict the values of ***SalePrice*** in the Train Data
Calculate the ***errors*** compared to actual
Tune parameters (***a***, ***b***) to minimize ***Cost***

Algorithmic Optimization

Hypothesize a Linear Model

$$\boldsymbol{SalePrice = a \times GrLivArea + b + \epsilon}$$

Cost Function to Minimize

$$J(a,b) = \sum (SalePrice - a \times GrLivArea - b)^2$$

$$J(a,b) = Residual\ sum\ of\ squares\ (RSS)$$

$$J(a,b) = n * (Mean\ square\ error(MSE))$$

Which parts of this Lesson were the hardest to grasp? You...

Connection between Correlation and Linear Regression	61 (11%)
The intuition of how well a Line fits, based on the Errors	56 (10%)
The algorithm to Minimize Cost Function in Regression	141 (25%)
The concept of Goodness of Fit and Mean Squared Error	102 (18%)
The concept of Explained Variance and R-Squared	196 (35%)

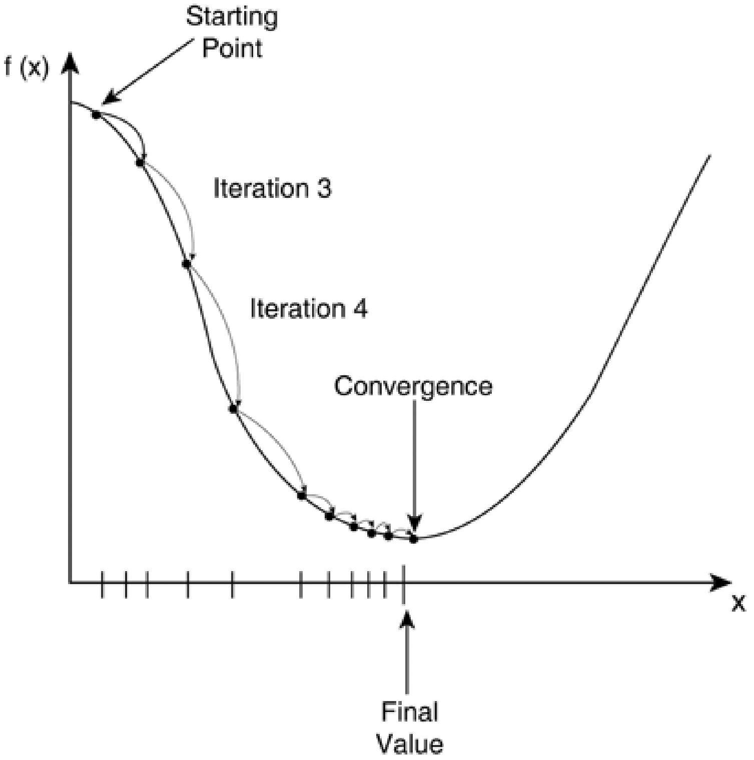
Which part of this Lesson will you like me to review in t...

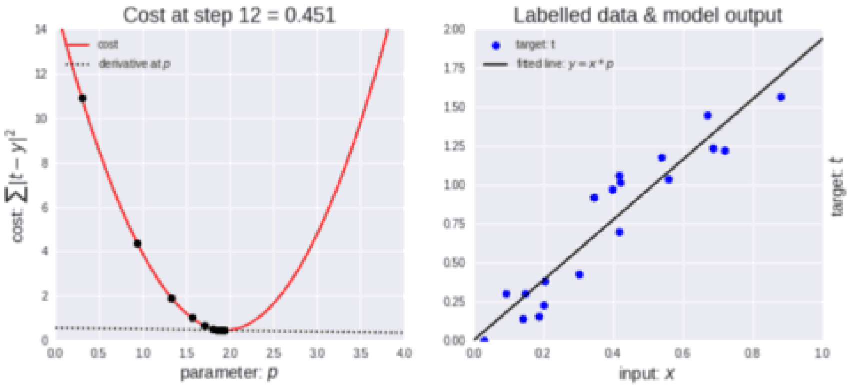
Connection between Correlation and Linear Regression	70 (11%)
The intuition of how well a Line fits, based on the Errors	67 (11%)
The algorithm to Minimize Cost Function in Regression	150 (24%)
The concept of Goodness of Fit and Mean Squared Error	125 (20%)
The concept of Explained Variance and R-Squared	211 (34%)

Data Science

Uni-Variate Regression

Gradient descent



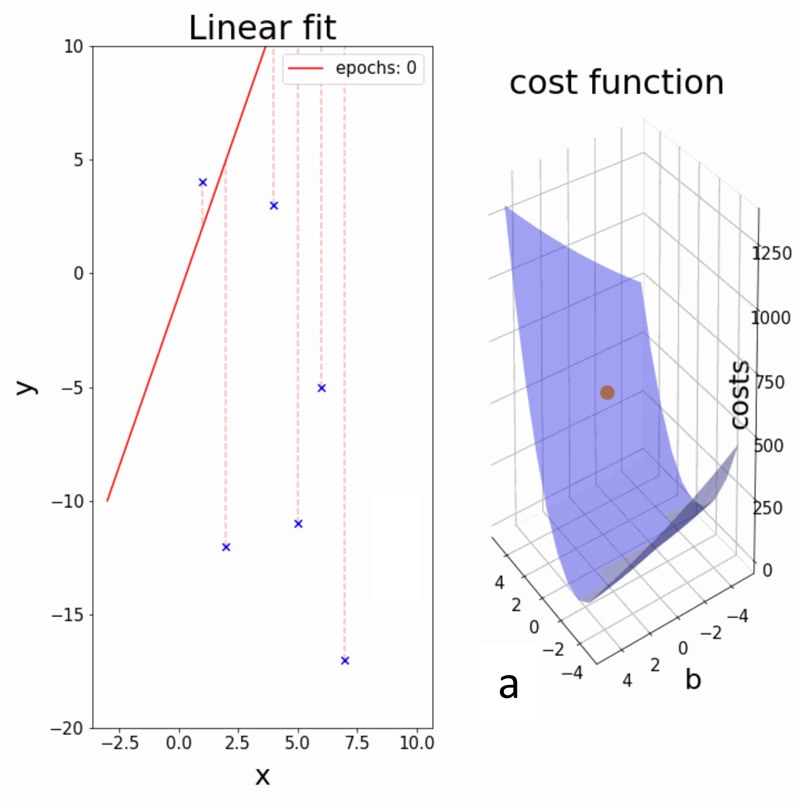


Do note that this graph is not depicting sale price and GrLivArea

Data Science

Uni-Variate Regression

Gradient descent



Do note that this graph is not depicting sale price and Grlivarea

Cost Function to Minimize

$$J(a,b) = \sum (SalePrice - a \times GrLivArea - b)^2$$

Minimizing Errors

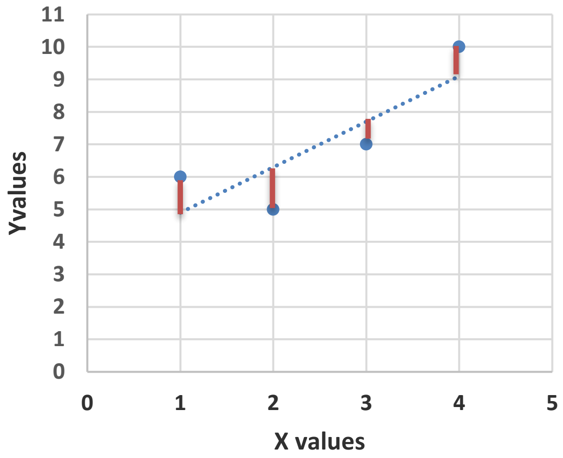
Analytic Solutions

Let's try Calculus first …

Regression Coefficients

X	Y	Ŷ
1	6	4.9
2	5	6.3
3	7	7.7
4	10	9.1

$\hat{Y}=aX+b$



Find the coefficients *a*, *b* **to minimize**

$$J(a,b) = \sum (y - ax - b)^2$$

Strategy : Set partial derivatives to zero.

$$\frac{\partial(J(a,b)}{\partial a} = -\sum 2x(y - ax - b)=0$$

-2[1(6-a-b)+ 2(5-2a-b)+3(7-3a-b)+4(10-4a-b)]=0
2(a+b-6+ (4a+2b-10)+ (9a+3b-21)+(16a+4b-40)) =0
60a+20b=154 --------(1)

$$\frac{\partial(J(a,b)}{\partial b} = -\sum 2(y - ax - b)=0$$

-2[(6-a-b)+ (5-2a-b)+(7-3a-b)+(10-4a-b)]=0
2(a+b-6+ (2a+b-5)+ (3a+b-7)+(4a+b-10)) =0
20a+8b=56 --------(2)

Solving a and b from (1) and (2)
a=1.4 and b=3.5

$\hat{Y}=1.4X+3.5$

X	Y	Y-Ybar	X-Xbar	(Y-Ybar)(X-Xbar)	$(X\text{-}Xbar)^2$
1	6	-1	-1.5	1.5	2.25
2	5	-2	-0.5	1	0.25
3	7	0	0.5	0	0.25
4	10	3	1.5	4.5	2.25

$\bar{x}$ =2.5 $\bar{y}$ =28/4=7

$\sum(y-\bar{y})(x-\bar{x})$ =7 **a=7/5=1.4**

$\sum(x-\bar{x})^2$ =5

$b=\bar{y}-a\bar{x}$ =7- 1.4*2.5=3.5 **b=3.5**

Ŷ=1.4X+3.5

Regression Coefficients

Find the coefficients *a*, *b* to minimize

$$J(a,b) = \sum(y - ax - b)^2$$

w.r.t. b : $b=\bar{y}-a\bar{x}$

w.r.t. a : $\dfrac{\sum(y-\bar{y})(x-\bar{x})}{\sum(x-\bar{x})^2}$

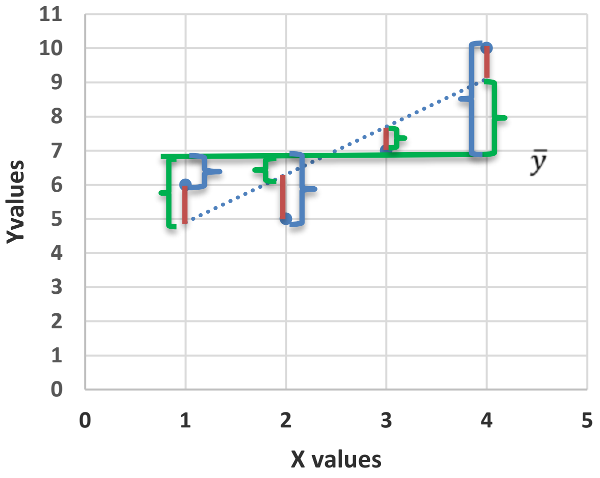
Visual Intuition : https://www.youtube.com/watch?v=3g-e2aiRfbU
Inspired by 3Blue 1Brown

X	Y	Ŷ	Y-Ŷ	(Y-Ŷ)^2	Y-Ybar	(Y-Ybar)^2	Ŷ-Ybar	(Ŷ-Ybar)^2
1	6	4.9	1.1	1.21	-1	1	-2.1	4.41
2	5	6.3	-1.3	1.69	-2	4	-0.7	0.49
3	7	7.7	-0.7	0.49	0	0	0.7	0.49
4	10	9.1	0.9	0.81	3	9	2.1	4.41

$\bar{y}$ =28/4=7

a=1.4 b=3.5

Ŷ=1.4X+3.5



To find best fit

R^2 :-how well the regression equation describes the relationship between the dependent variable (*Y*) and the independent variable (*X*).

$$J(a,b) = \sum (y - ax - b)^2$$

RSS=$\sum(Y - \hat{Y})^2$ =4.2 **Un explained variation= error**

Benchmark design is when $\hat{Y} = \bar{y}$ **a=0, b=ybar=7**

Ŷ=0*X+7

TSS=$\sum(Y - \text{Ybar})^2$=14

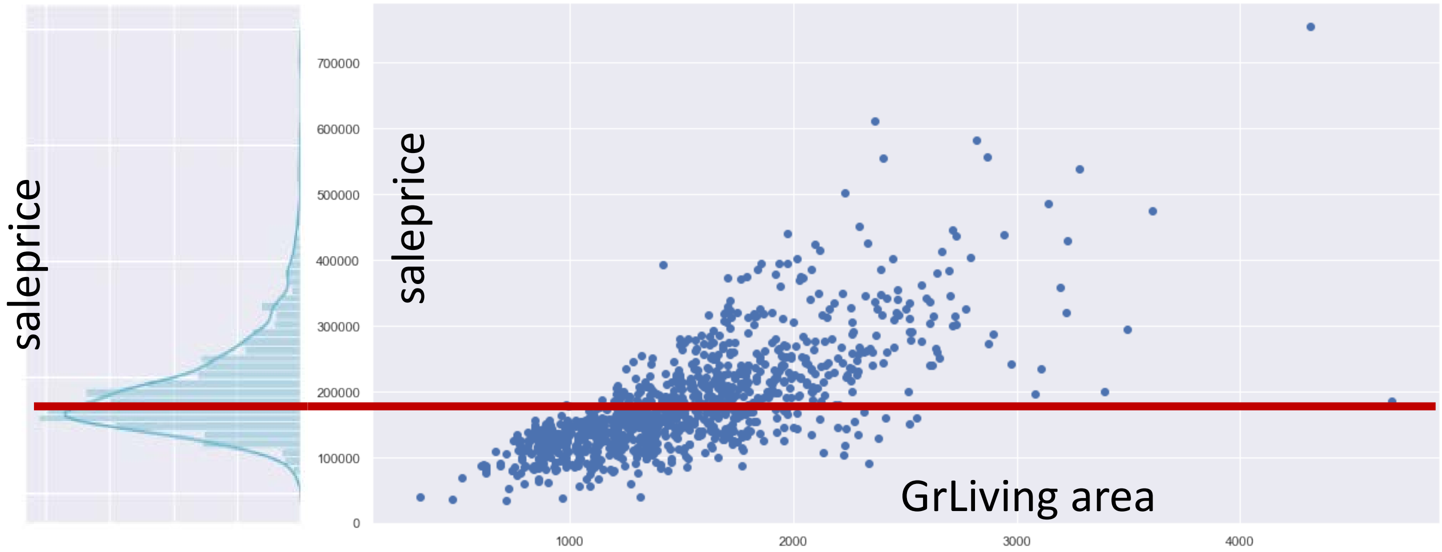
Explained sum of squares (ESS)=$\sum(\hat{Y} - Ybar)^2$

ESS=$\sum(\hat{Y} - Ybar)^2$=9.8 **Explained variation**

TSS= RSS+ESS= 9.8+4.2=14

R^2= (TSS-RSS)/TSS = ESS/TSS= 9.8/14=0.7

Benchmark design



Mean = 181253

Var = 6333613056

Std.Dev = 79584

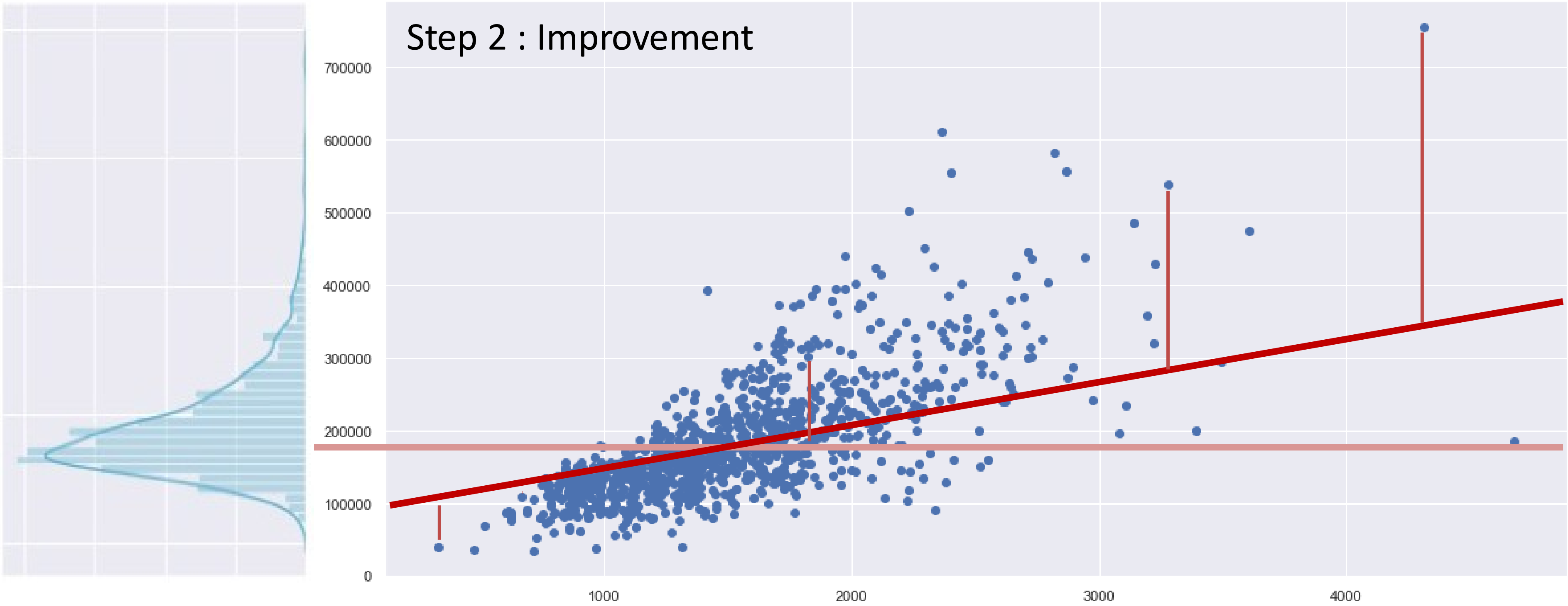
Model = 181253 + 0 x GrLivArea

Mean Squared Error (MSE) = 6333613056

Root Mean Squared Error (RMSE) = 79584

(a, b) = (0, 181253)

$R^2 = 0$



Mean = 181253

Var = 6333613056

Std.Dev = 79584

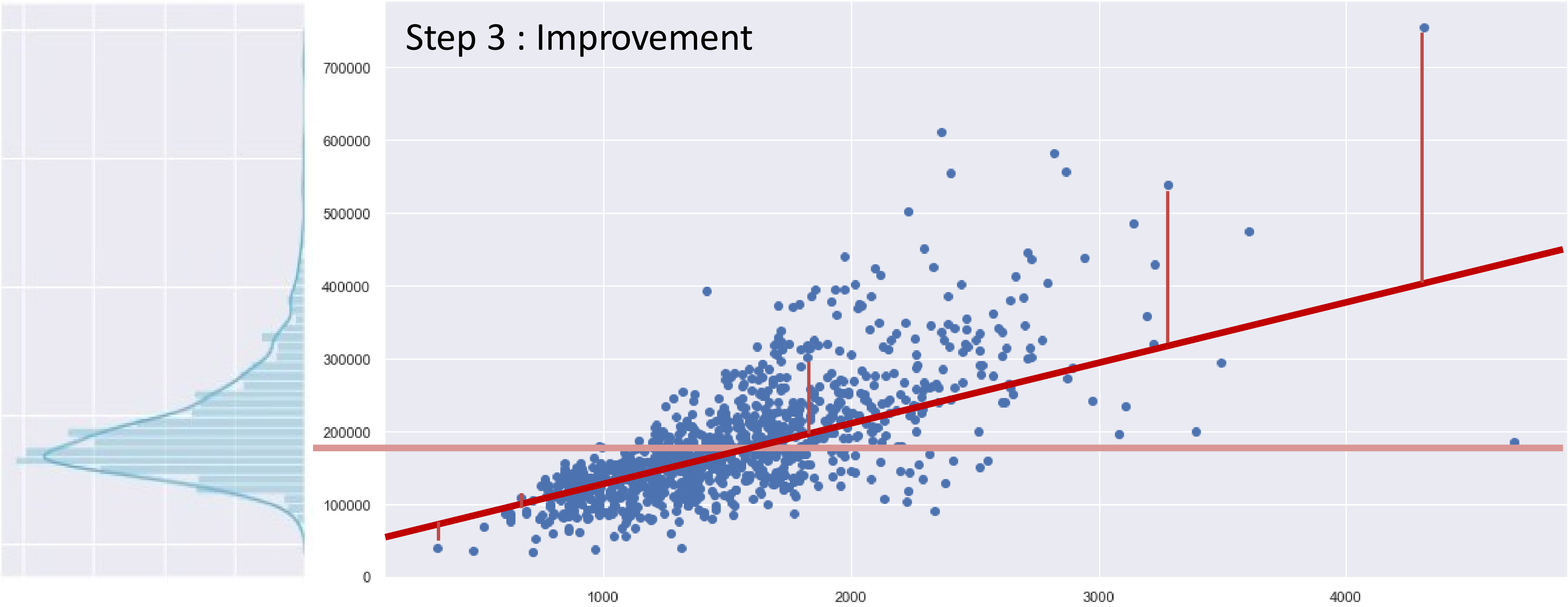
Model : y = 21 . X + 106710

TSS = n x 6333613056 = n x Var

RSS = n x 5573579489 = n x MSE

(a, b) = (21, 106710)

$R^2 = 0.12$



Mean = 181253

Var = 6333613056

Std.Dev = 79584

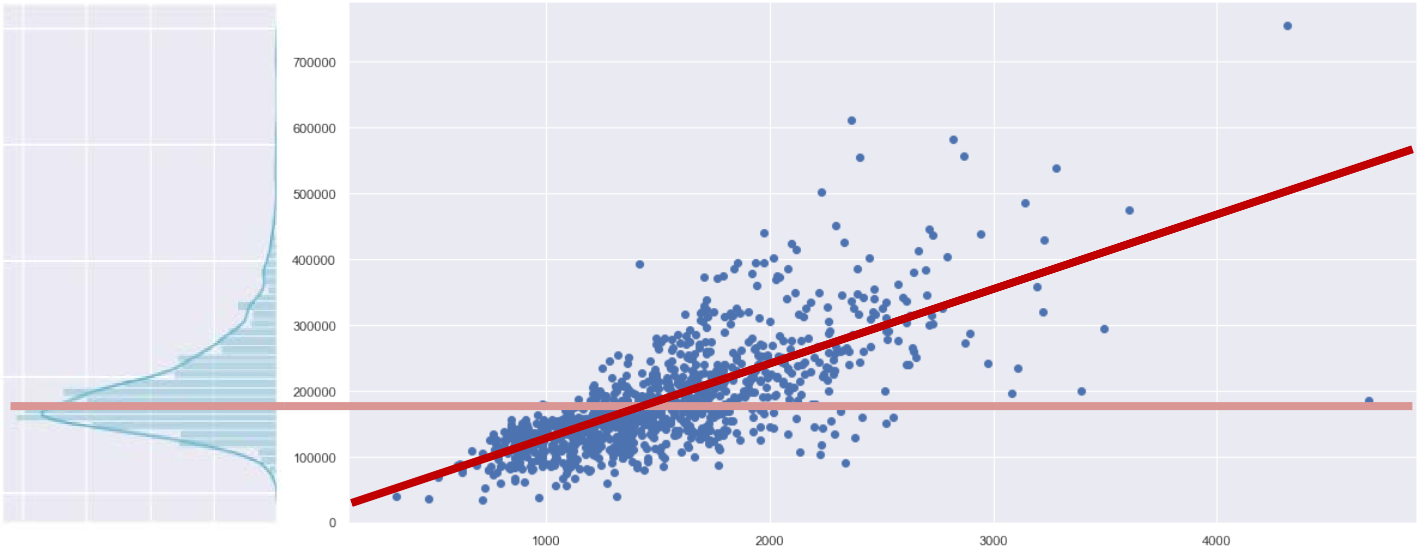
Model : y = 78 . X + 58923

TSS = n x 6333613056 = n x Var

RSS = n x 3483487180 = n x MSE

(a, b) = (78, 58923)

$R^2 = 0.45$



Mean = 181253

Var = 6333613056

Std.Dev = 79584

Model = 9498 + 113 x GrLivArea

Mean Squared Error (MSE) = 2976798136

Root Mean Squared Error (RMSE) = 54560

(a, b) = (113, 9498)

$R^2 = 0.53$

Linear Regression

Benchmark: Y = 181253 + 0 *X

Estimate:- Mean of y(train data)

Mean square error(MSE)= variance of y in train

$1/n[(Y-\hat{Y})^2]=1/n[(Y-\text{Ybar})^2]$

RSS= TSS

$R^2 = 1 - \frac{RSS}{TSS}$ $R^2 = \frac{TSS - RSS}{TSS}$ $R^2 = \frac{Var - MSE}{Var}$

Explained variance or coefficient of determination R^2 measures how good is your final model compared to your bench mark.

Final Model: Y = 9498 + 113 * X

Estimate:- Value of y using model(train data)

Mean square error(MSE)< variance of y in train

RSS=TSS-ESS= n*MSE

Any relation to “Correlation”?

Linear Regression

Important things to remember

1. Check weather the regression dataset Columns/Features as are Linear or Non-Linear (has a skew) as RMSE would be pretty high for Non-Linear data.
2. Removing or Imputing outliers with mean or median might help reduce errors . Box-plots for each column might help find outliers.
3. If the features are Non-Linear apply logarithmic , square , cubic to the column to make it Standard Normalized Distribution to proceed further for Machine Learning model