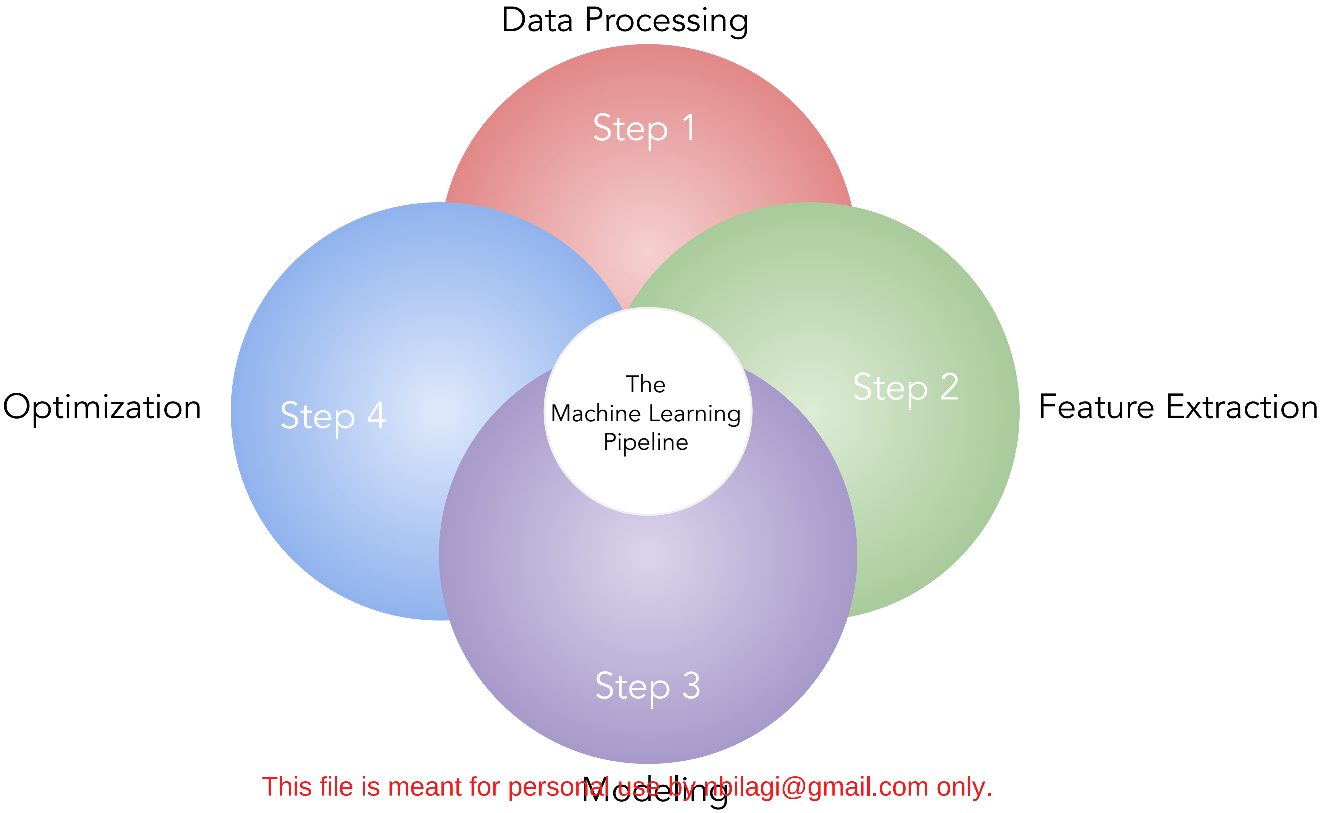
Feature Engineering & Feature Selection

Agenda

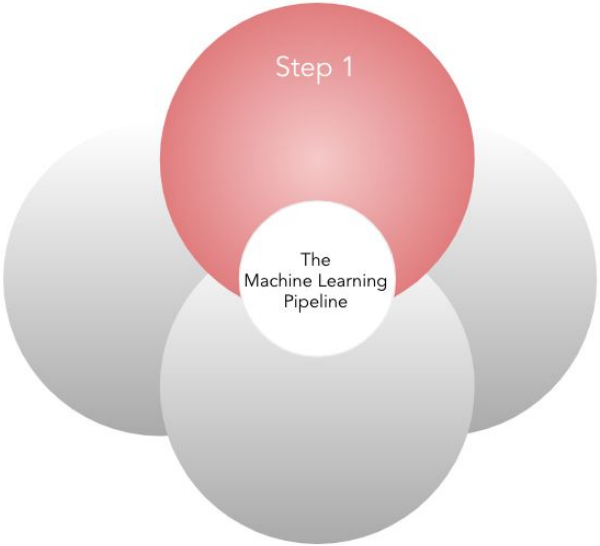
- Feature Transformation
- Feature Scaling
- Feature Selection
  - Forward Selection
  - Backward elimination
  - Stepwise selection
  - Recursive Feature Elimination (RFE)

Machine Learning Pipeline

The ML pipeline



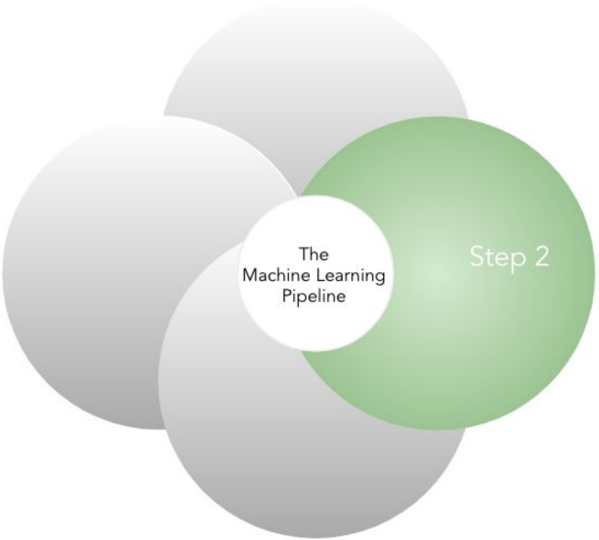
The ML pipeline: Data processing



DATA PROCESSING

- Collection
- Formatting
- Labelling

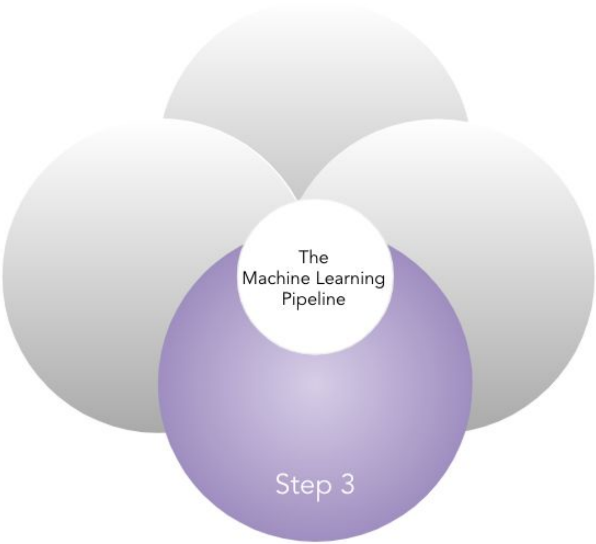
The ML pipeline: Feature extraction



FEATURE EXTRACTION

- Feature Transformation
- Feature Engineering
- Feature Selection

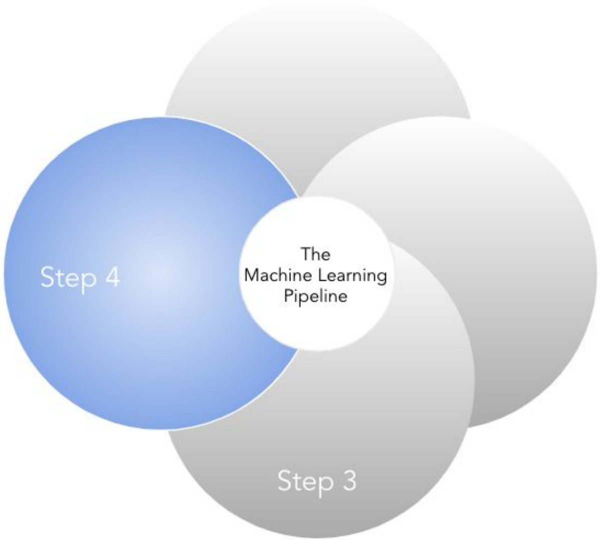
The ML pipeline: Modeling



MODELING

- Model Building
- Model Evaluation

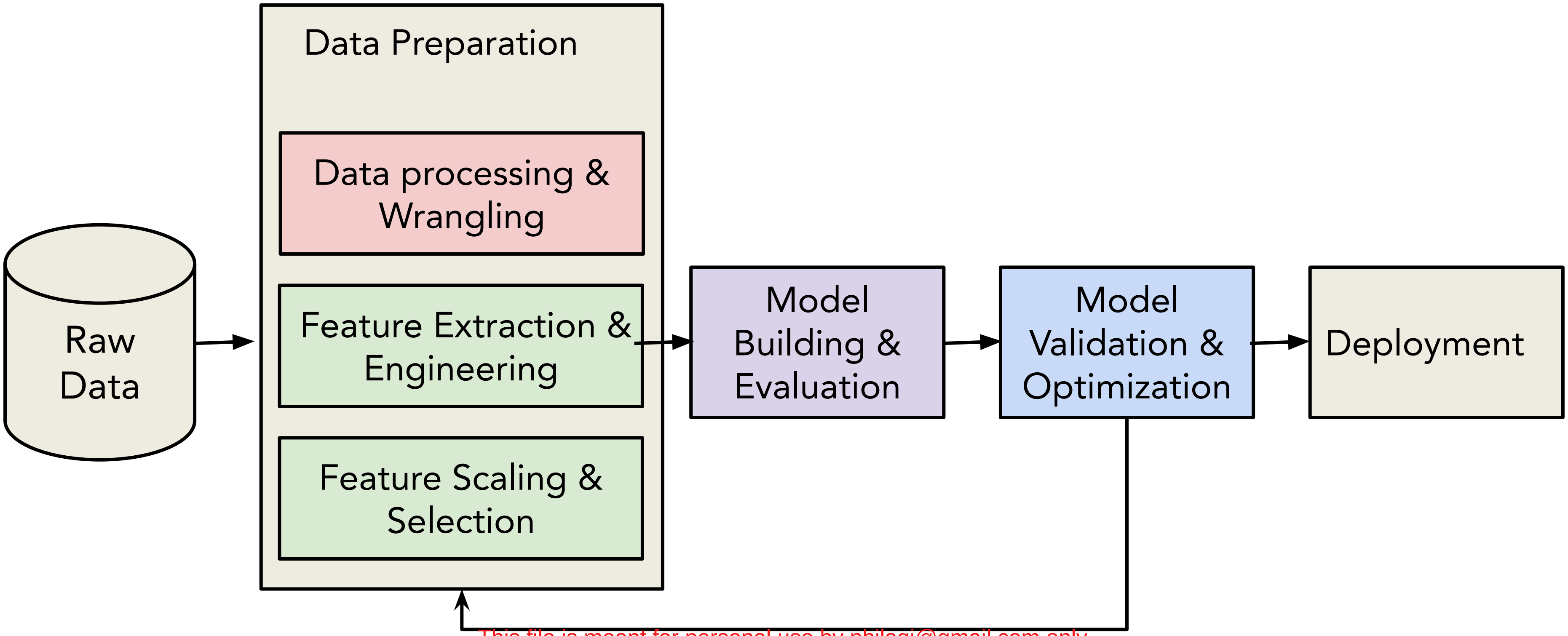
The ML pipeline: Optimization



OPTIMIZATION

- Prediction Evaluation
- Model Validation
- Fine Tuning

The ML pipeline



Data Processing

Data processing

DATA PROCESSING

- Collection
- Formatting
- Labelling

- Collection: To extract data from various sources. Generally obtained in the raw form and not immediately suitable for analysis
- Formatting: Organizing the datasets as required for analysis
- Labelling: Manually labelling data

Feature Extraction

Feature

- Feature or attribute is an independent variable that acts as input to our model
- The columns of a dataset are considered as features

Features

Product_ID	Store	City
FD_234	A	Chennai
DR_543	A	Bangalore
FD_176	B	Mumbai
DR_621	A	New Delhi

Feature Extraction

FEATURE EXTRACTION

- Feature Transformation
- Feature Engineering
- Feature Selection

- Feature Transformation: Replacing the existing features by function of these feature
- Feature Engineering: Creating new features based on empirical relationships
- Feature Selection: Fitting a model of significant features

Feature Transformation

Why do we need feature transformation?

- Incase of skewed (predictor and/or dependent) variable, we transform it to reduce the skewness

- If the assumptions of linear regression are not met, transformation of skewed target variable can be used for making the error terms more compatible to the assumptions

- If the relationship between a predictor and the response variable is non-linear, it can be linearized using transformation

Please!! Note

Assumption of normality

The parametric methods used to compute test statistics or confidence intervals on the predictor variables assume the data to follow a normal distribution

Hence it is favourable that features have approximately normal distribution

Recap: The parametric methods are used when sample statistics adequately represent the population

Rule for transformed variables

Comparison of model performance should be done using the original units for the target variable and not the units after transformation

Transformation methods

- Logarithmic transformation
- Square root transformation
- Reciprocal transformation
- Exponential transformation
- Box-cox transformation

Transformation methods

- Logarithmic transformation
- Square root transformation
- Reciprocal transformation
- Exponential transformation
- Box-cox transformation

Logarithmic transformation

- To linearize, values of a variable are replaced with its natural log
- It cannot be used on a categorical variable after dummy encoding since ln(0) is undefined
- Also if a variable takes zero or negative values, logarithmic transform cannot be used on it

Example of log transformation

- If the relationship between x and y is given by:

$$y = mx^k\varepsilon$$

To transform to a linear relationship take logarithm on both sides:

$$\ln(y) = \ln(m) + k * \ln(x) + \ln(\varepsilon)$$

$$Y = \beta_0 + \beta_1 X + \varepsilon'$$

Y	$\ln(y)$
β_0	$\ln(m)$
β_1	k
X	$\ln(x)$
ε'	ε

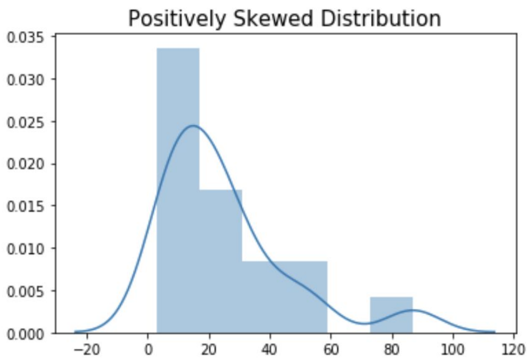
- Now a regression line can be estimated for this relationship

Example of log transformation

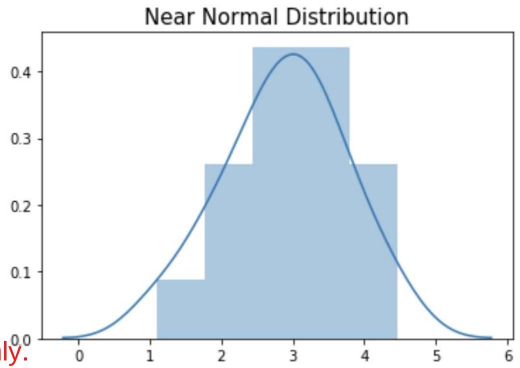
Consider the following data:

Note: Values are rounded to 1 decimals

X	12	9	3	6	24	13	21	6	16	13	54	23	46	32	87	23	34
ln(X)	2.5	2.2	1.1	1.8	3.2	2.6	3.1	1.8	2.7	2.6	3.9	3.1	3.8	3.4	4.5	3.1	3.5



Logarithmic Transformation



Transformation techniques

- Logarithmic transformation
- Square root transformation
- Reciprocal transformation
- Exponential transformation
- Box-cox transformation

Square root transformation

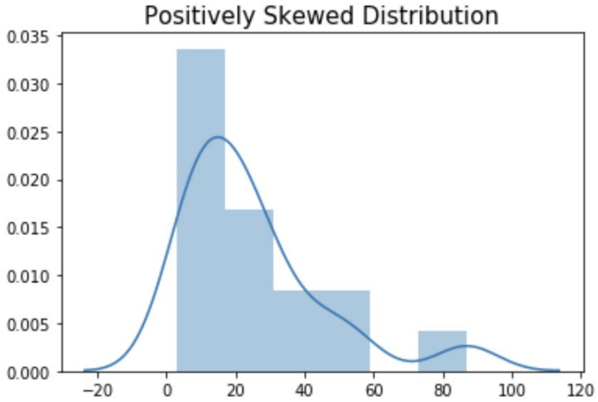
- Values of a variable are replaced with its square root
- To reduce right skewness, we may use square root transformation
- It can be applied even when the variable takes a zero value

Example of square root transformation

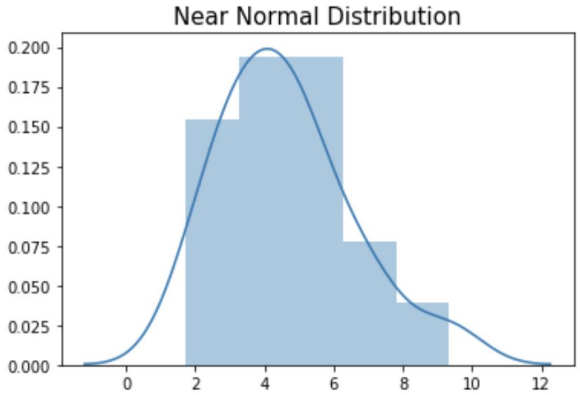
Consider the following data:

Note: Values are rounded to 1 decimals

X	12	9	3	6	24	13	21	6	16	13	54	23	46	32	87	23	34
$\sqrt{X}$	3.5	3	1.7	2.4	4.9	3.6	4.6	2.4	4	3.6	7.4	4.8	6.8	5.7	9.3	4.8	5.8



Square Root Transformation



Transformation techniques

- Logarithmic transformation
- Square root transformation
- Reciprocal transformation
- Exponential transformation
- Box-cox transformation

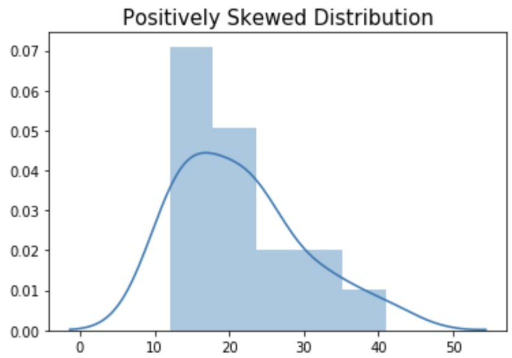
Reciprocal transformation

- Values of a variable are replaced with its reciprocal
- It can not be applied only when the variable takes zero values
- However, can be applied to negative values
- Example: population per area (population density) transforms to area per person

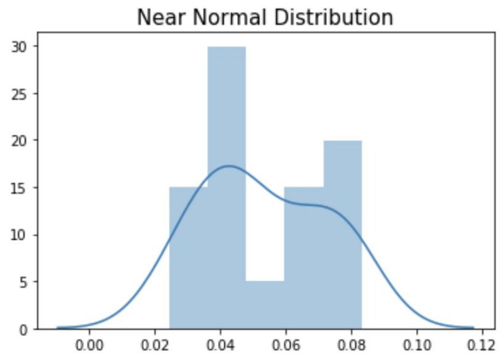
Example of reciprocal

Consider the following data :

X	12	19	23	16	14	13	21	13	16	13	24	23	41	32	27	23	34
1/X	.08	.05	.04	.06	.07	.08	.05	.08	.06	.08	.04	.04	.02	.03	.04	.04	.03



Reciprocal Transformation



Transformation techniques

- Logarithmic transformation
- Square root transformation
- Reciprocal transformation
- Exponential transformation
- Box-cox transformation

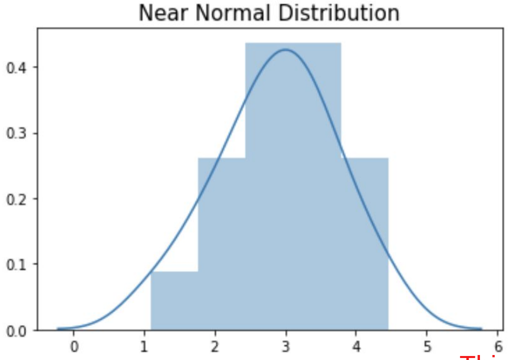
Exponential transformation

- Values of a variable are replaced with its exponential
- It is generally used to transform logarithmic transformed data to get the original data back

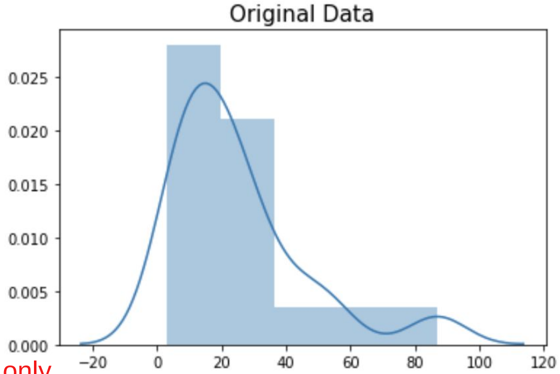
Example of exponential transformation

Consider the data used in logarithmic transformation.

X	12	9	3	6	24	13	21	6	16	13	54	23	46	32	87	23	34
ln(X)	2.5	2.2	1.1	1.8	3.2	2.6	3.1	1.8	2.7	2.6	3.9	3.1	3.8	3.4	4.5	3.1	3.5
exp(X)	12	9	3	6	24	13	21	6	16	13	54	23	46	32	87	23	34



Exponential Transformation



Transformation techniques

- Logarithmic transformation
- Square root transformation
- Reciprocal transformation
- Exponential transformation
- Box-cox transformation

Box cox transformation

- It is defined as

$$X^{\lambda} = \begin{cases} \frac{X^{\lambda}-1}{\lambda} & \text{if} \lambda > 0 \\ \ln(X) & \text{if} \lambda = 0 \end{cases}$$

Here, X is the variable and λ is the transformation parameter and can be tuned according to the data.

- The Box-Cox transformation can only be used on positive variables
- Generalized form of logarithmic transformation

Feature Scaling

Feature scaling

- It is a technique used to transform the data into a common scale

- Since the features have various ranges, it becomes a necessary step in data preprocessing while using machine learning algorithms

- Since most machine learning algorithms use distance calculations, features taking higher values will weigh in more in the distance compared to features taking values of low magnitude

Example

- In a dataset which has variables age and income. The age of a person is measured in years which can takes values between 18 to 65 (retirement age) and income of a person is in thousands
  So it is necessary to bring the two features in the same scale to assign appropriate weights

- In some parts of the world height is measured using metric system (centimetres), while in some other parts the imperial system is used (feet/inches).
  So the results would be different if the height value is 152 cm or 5 feet, when if converted they refer to the exact same height value.

Feature scaling methods

- Normalization
- Standardization

Normalization

Normalization is the process of rescaling features in the range 0 to 1

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Standardization

- Standardization rescales the feature such that it has mean 0 and unit variance
- The procedure involves subtracting the mean from observation and then dividing by the standard deviation

$$x' = \frac{x - \bar{x}}{\sigma}$$

When to use Normalization? When to use Standardization?

Normalization is a good technique to use when you do not know the distribution of your data or when you know the distribution is not Gaussian (a bell curve).

Standardization assumes that your data has a Gaussian (bell curve) distribution. This does not strictly have to be true, but the technique is more effective if your attribute distribution is Gaussian.

Normalization or Min-Max normalization tries to get the values closer to mean, but when there are outliers in the data which are important and we don't want to lose their impact, we go with Standardization or Z score normalization

Min- Max tries to get the values closer to mean. But when there are outliers in the data which are important and we don't want to lose their impact, we go with Z score. In this case, we rescale an original variable to have a mean of zero and a standard deviation of one. It does not have any units: hence is useful for comparing variables expressed in different units.   Standardization makes no difference to the shape of a distribution.

Feature Selection

Feature selection

- Feature selection is the process of including the significant features in the model
- This can be achieved by:
    - Forward selection method
    - Backward elimination method
    - Stepwise method
- To understand the above methods let $X_1, X_2, \ldots, X_k$ be k predictor variables and Y be the response variable

Forward selection method

Procedure

1. Start with a null model (with no predictors)

2. Obtain the correlation between Y and each variable. The variable with highest correlation gets added to the model (say X_m). Build a model $Y \sim X_m$

3. Obtain the correlation between Y and remaining (k-1) variables. The next variable (say X_p) is included, which has the highest correlation with Y after removing X_m

4. Build a model $Y \sim X_m + X_p$. If X_p is significant include it in the model else discard

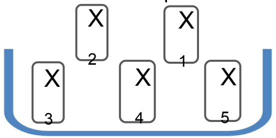
5. Repeat steps (3) and (4) until reaching the stopping rule or running out of variables

Forward selection method

Start with a NULL MODEL
(a model with no predictors)

$Y \sim$

Consider 5 predictors

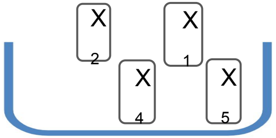


Obtain the most significant predictor
(predictor having highest correlation with Y)

Model with most significant variable
(say X_3)

$Y \sim \beta_0 + \beta_1 X_3$

Add to the model

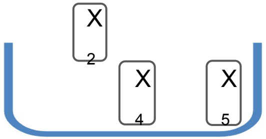


Obtain the next most significant predictor
(from the remaining 4 predictor)

Model with most significant variable
(say X_1)

$Y \sim \beta_0 + \beta_1 X_3 + \beta_2 X_1$

Add to the model



Continue until reaching the stopping rule or running out of variables

Backward elimination method

Procedure

1. Start with a full model (model with all k predictors)
2. Remove the variable which is least significant (variable with largest p-value)
3. Fit a new model with remaining (k-1) regressors
4. The next variable (say X_p) is removed if it is least significant
5. Repeat steps (3) and (4) until reaching the stopping rule or all variables are significant

Backward elimination method

Start with a FULL MODEL
(a model with all the 5 predictors)

$$Y \sim \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \beta_5X_5$$

Obtain the least significant predictor
(predictor having highest p-value)

Model after removing the least significant variable
(say X_3 the least significant)

Remove X_3

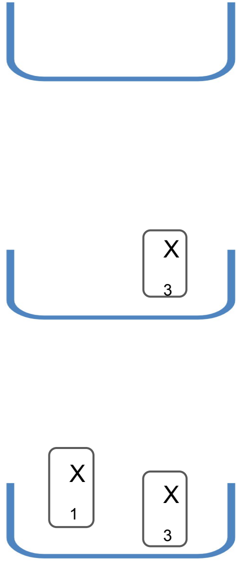
$$Y \sim \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_4 + \beta_4X_5$$

Obtain the next least significant predictor
(predictor having highest p-value after removing X_3)

Model after removing the least significant variable
(say X_1 is least significant)

Remove X_1

$$Y \sim \beta_0 + \beta_1X_2 + \beta_2X_4 + \beta_3X_5$$

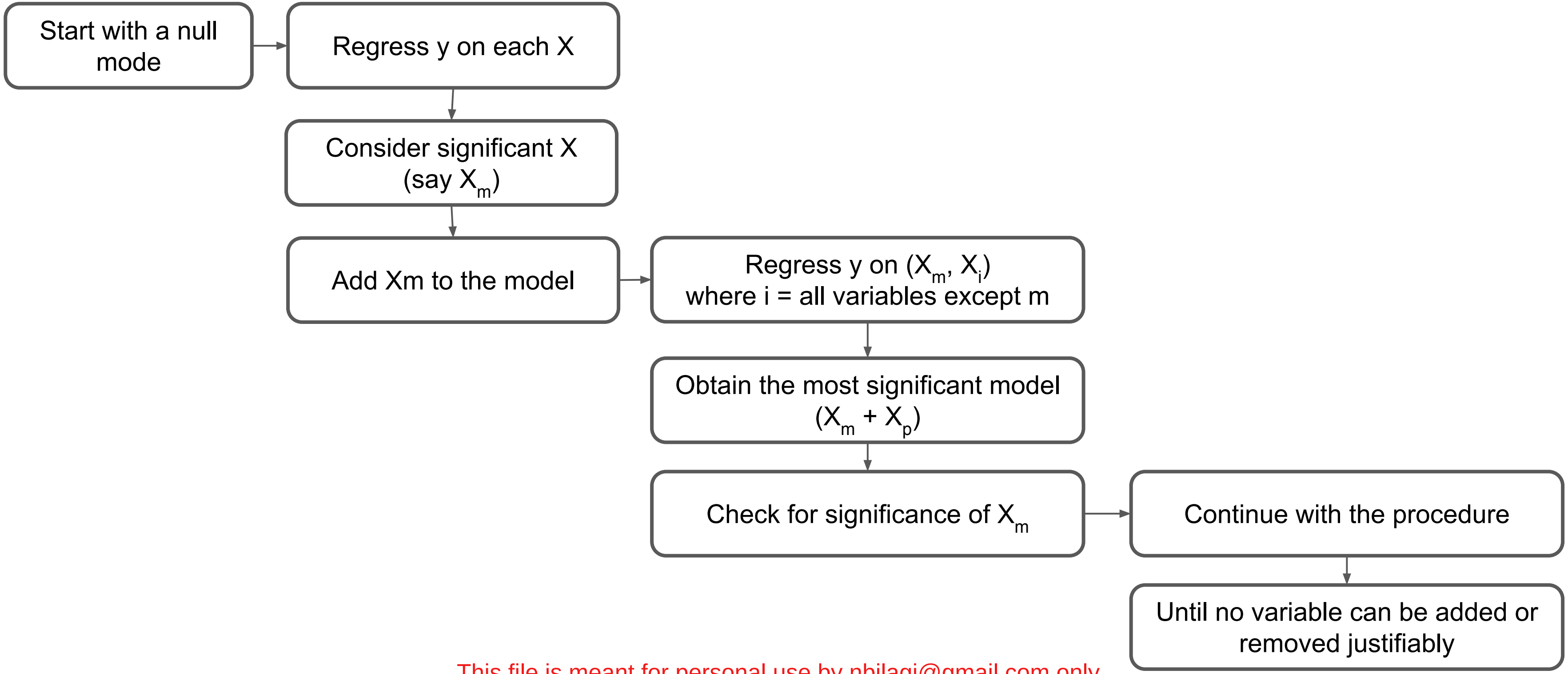


Continue until reaching the stopping rule or running out of variables

Stepwise regression

- It is a combination of forward selection and backward elimination method
- Procedure:
  - Start with a null model (with no predictors)
  - At each step add or remove variable based on its corresponding p-value
  - Stop when no variable can be added or removed justifiably

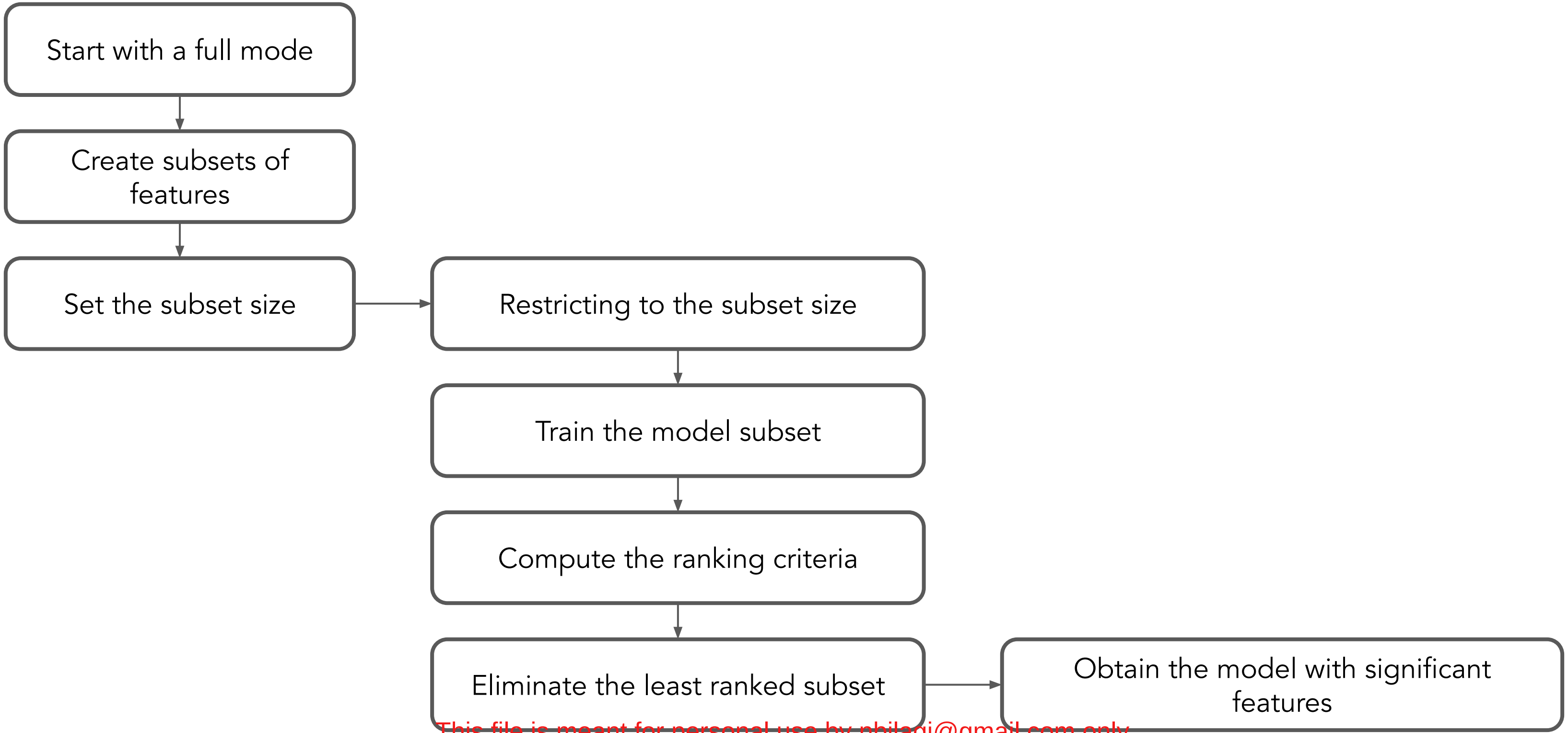
Stepwise regression



Recursive feature elimination (RFE)

- It is an instance of backward feature elimination
- Procedure:
  - Train a full model
  - Create subsets for features
  - Set the subset size
  - Compute the ranking criteria for each feature subset
  - Remove the feature subset that has the least ranking

Recursive feature elimination (RFE)



Recursive feature elimination (RFE)

Algorithm 1: Recursive feature elimination

```
Tune/train the model on the training set using all predictors
Calculate model performance
Calculate variable importance or rankings
for Each subset size S_i, i = 1 ... S do
    Keep the S_i most important variables
    [Optional] Pre-process the data
    Tune/train the model on the training set using S_i predictors
    Calculate model performance
    [Optional] Recalculate the rankings for each predictor
end
Calculate the performance profile over the S_i
Determine the appropriate number of predictors
Use the model corresponding to the optimal S_i
```