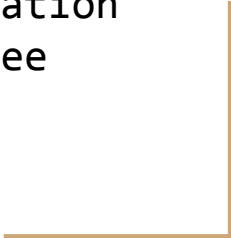




U:Pass Week 3

36106: Machine Learning
Algorithms and Application
By Marisara Satrulee



Can Do

- 20 minutes of discussion on the key points of session for each week
- 40 minutes of coding exercise and/or discussion activities
- Take questions and discuss problem in class

*** I am **not** a subject-matter expert, for questions I cannot answer I will get back to you in the following session ***

Can't Do

- Cannot give you my personal contact

For any studying assistant please
email helps@uts.edu.au or
call 02 9514 9733

- Discuss assignments
-

Recap of week 1-2

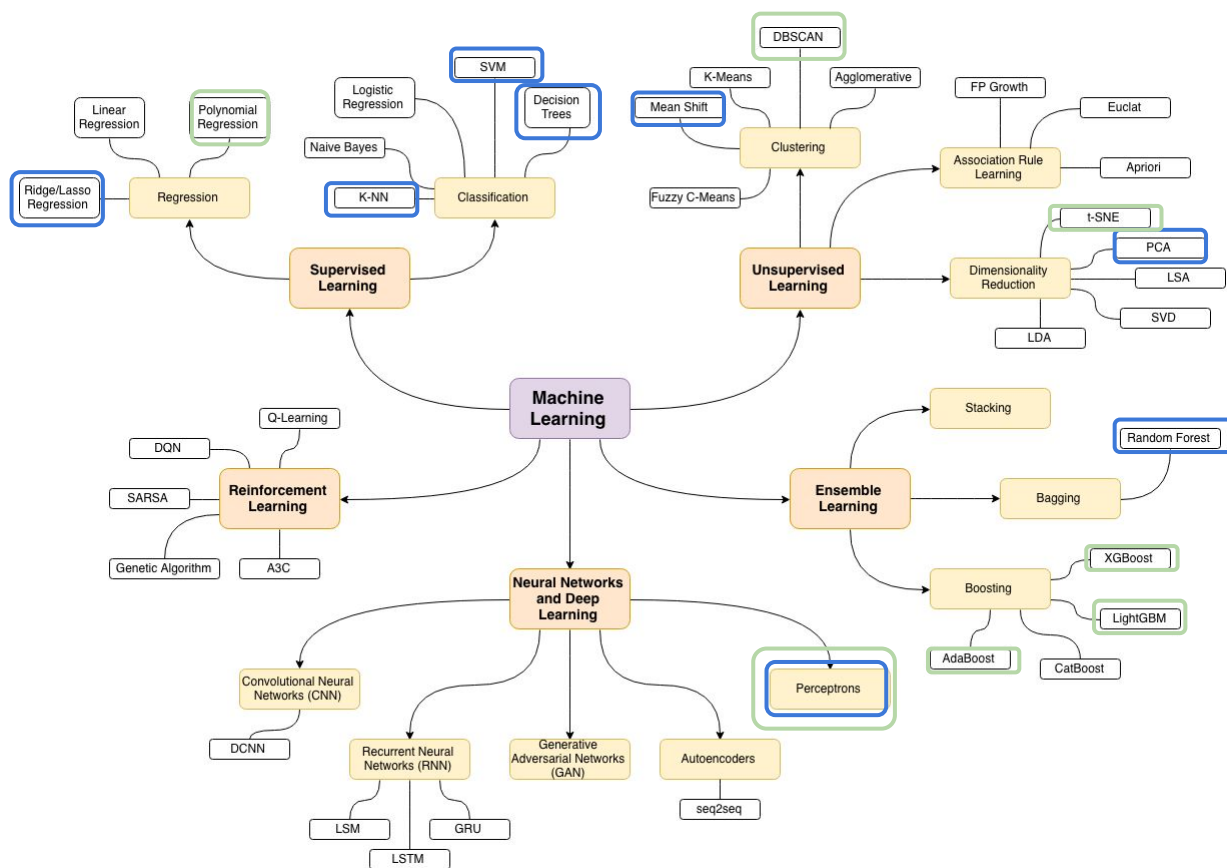
Week 1

- CRISP-DM
- Regression Model
- Parameters
 - Model Parameters vs Hyperparameters

Week 2

- Performance Metrics
- Data Splitting
- Regularization










ML Models

 36106

 36120

- [Ensemble](#)
- [Reinforcement Learning](#)

How to Choose ML Algorithm

	Input (Data)
	Output (Goal)
	Field Of Study
	Limitations
	Preferences

1. Define the Problem
2. Understand the Data
3. Consider Model Complexity
4. Evaluate Performance Metrics
5. Cross-Validation and Model Evaluation
6. Experiment with Multiple Models
7. Regularization and Hyperparameter Tuning
8. Consider Model Explainability and Business Constraints

Today

1,2,4,
7,8

1. Define the Problem

Problem A:

Predict the student's test score on a scale of 0 - 100.

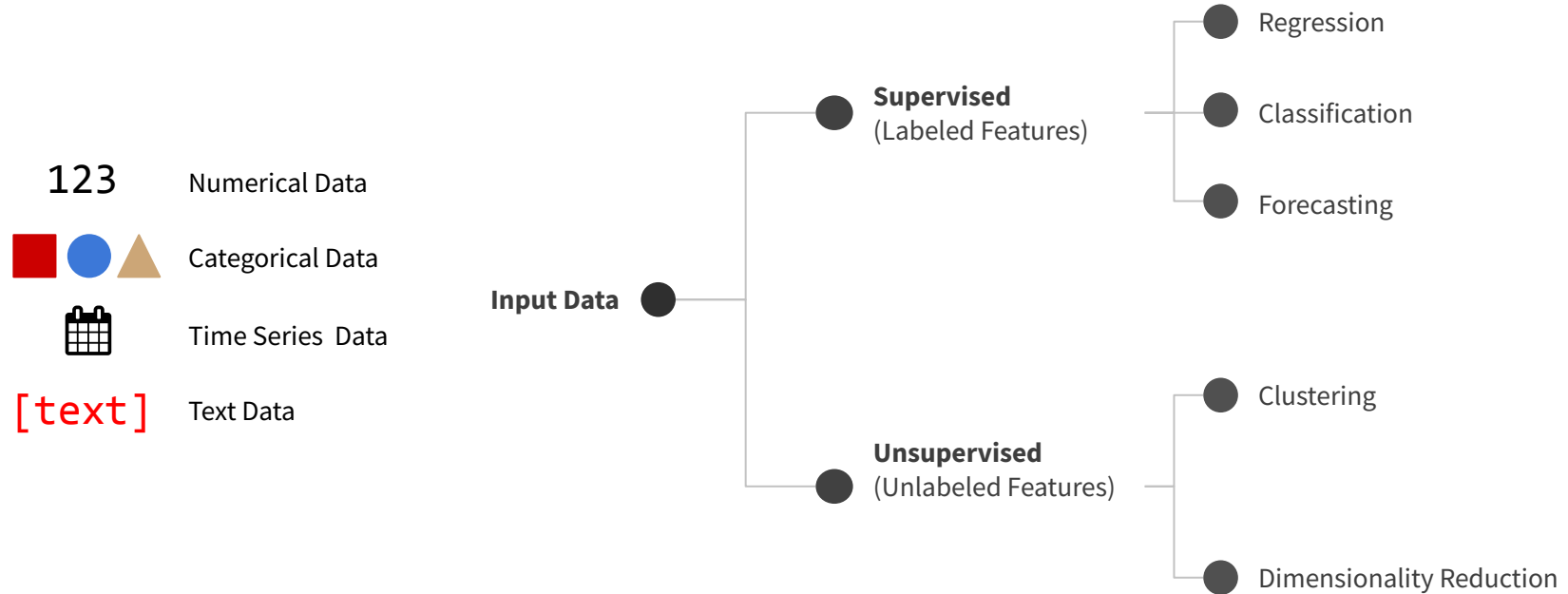
*** Predictions are continuous (numbers in a range).

Problem B:

Predict whether the student passed or failed.

*** Predictions are discrete (only specific values or categories are allowed)

2. Understand the Data



Type of Regression (seen so far)

Regression analysis estimates the relationship between two or more variables.

1. Linear Regression
2. Logistic Regression
3. Ridge Regression
4. Lasso Regression
5. Elastic Net Regression

Use when data shows **multicollinearity**, or the independent variables (input features) are highly correlated.

For example of Linear Regression:

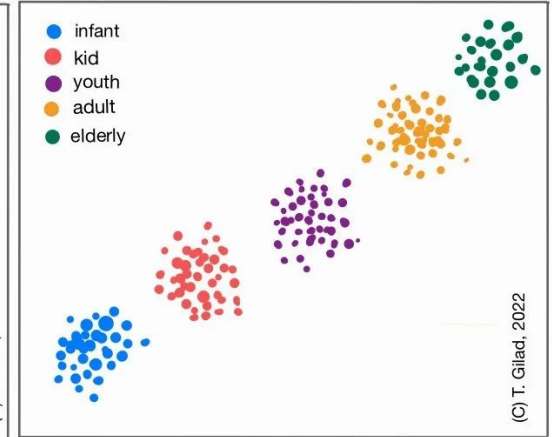
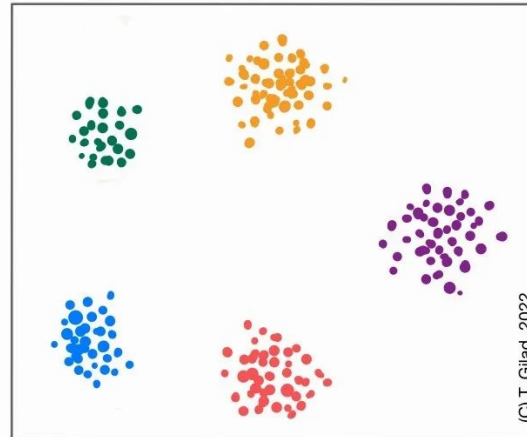
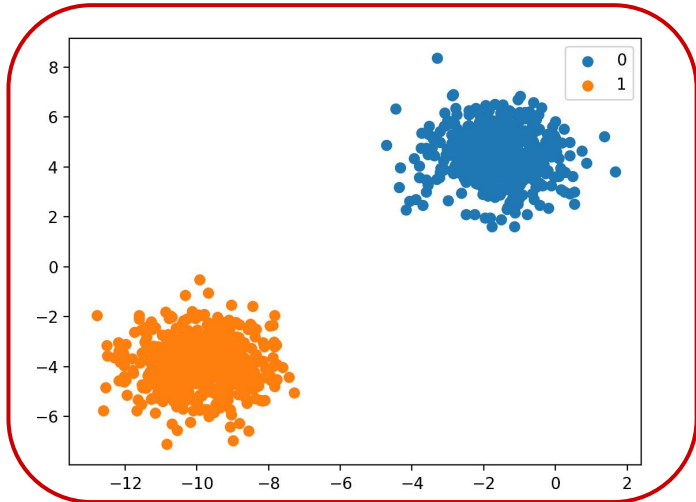
The relationship between the total number of hours studied and total prep exams taken on predicting students' final exam scores.

	Study Hours	Prep Exams	Final Exam Score
Student 1	3	2	76
Student 2	7	6	88
Student 3	16	5	96
Student 4	14	2	90
Student 5	12	7	98
Student 6	7	4	80
Student 7	4	4	86
Student 8	19	2	89
Student 9	4	8	68
Student 10	8	4	75
Student 11	8	1	72
Student 12	3	3	76

Type of Classification

Types of logistic regression

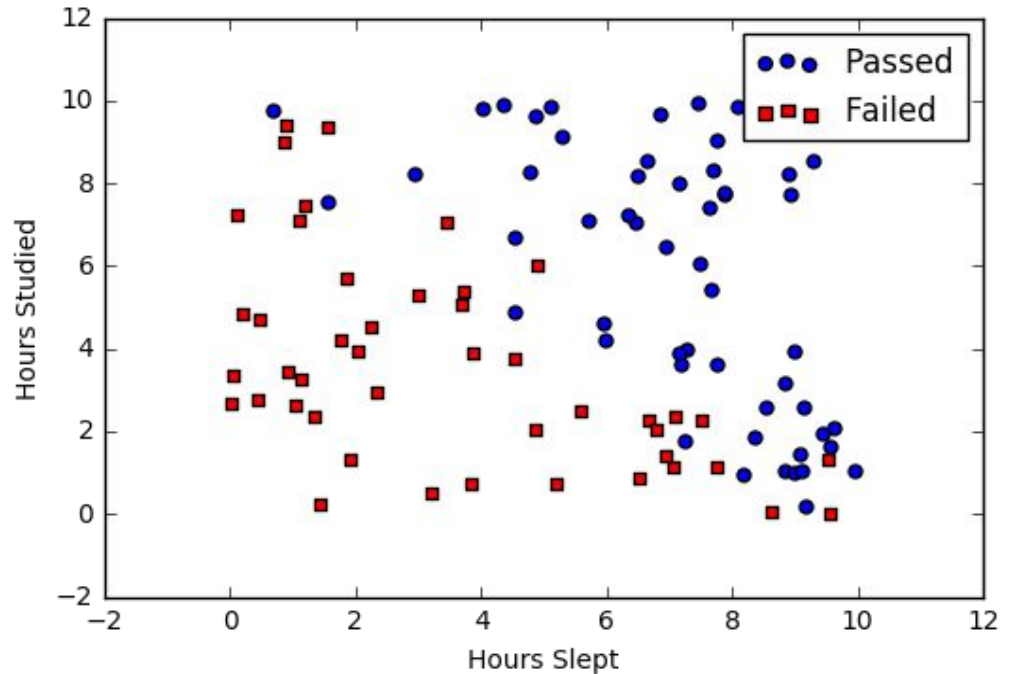
- Binary (Pass/Fail)
- Multi (Cats, Dogs, Sheep)
- Ordinal (Low, Medium, High)



Binary Classification

Example of **Binary logistic regression**

Studied	Slept	Passed
4.85	9.63	1
8.62	3.23	0
5.43	8.23	1
9.21	6.34	0



4. Evaluate Performance Metrics

- Mean Square Error (MSE)

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

- Root Mean Square Error (RMSE)

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y})^2}$$

- Mean Absolute Error (MAE)

$$\text{MAE} = \frac{1}{n} \sum_{j=1}^n |y_j - \hat{y}_j|$$

- R Square

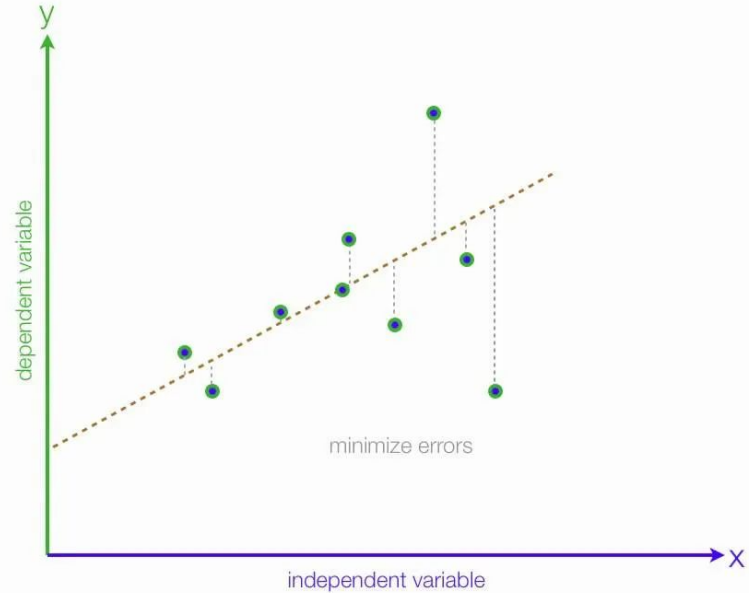
$$\hat{R}^2 = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

4. Evaluate Performance Metrics

The objective of Linear Regression is to find **a line** that **minimizes** the **prediction error** of all the data points.

The **lower** value of MAE, MSE, and RMSE implies **higher accuracy** of a regression model.

But the **higher** value of **R Square** is not always necessary indicating a good-fitted model.



4. Evaluate Performance Metrics

Both RMSE and R - Squared **quantifies how well** a linear regression model fits a dataset.

The **RMSE** tells how good a regression **model** can **predict** the value of a response variable (target) while **R - Squared** tells how well the predictor **variables** (input features) can **explain** the variation in the response variable (Chugh, 2024).

Typically, **R - Squared** is not suitable to judge a model, R^2 should be used: only as a measure of correlation between two sequences. (Dunn, 2022).

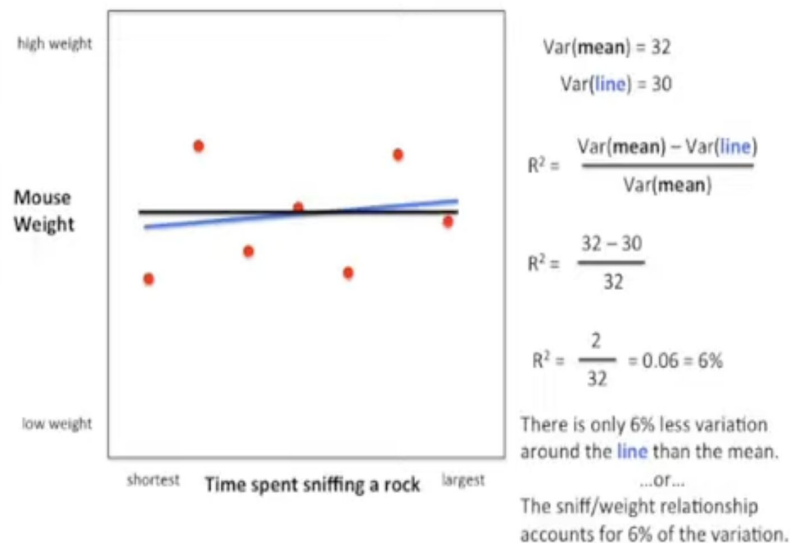
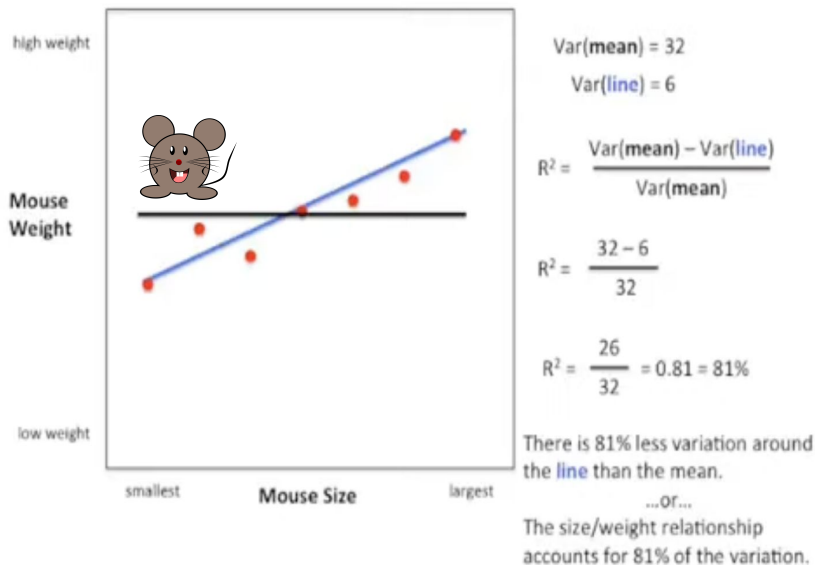
Reference:

Chugh, A. (2024, January 18). MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better? *Medium*.
<https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e>

Dunn, K. (2022, March 11). Avoid R-squared to judge regression model performance. *Medium*. <https://towardsdatascience.com/avoid-r-squared-to-judge-regression-model-performance-5c2bc53c8e2e>

4. Evaluate Performance Metrics

Experiment on features that contribute to a mouse's weight.



Reference:

<https://www.youtube.com/watch?v=bMccdk8EdGo>

7. Regularization & Hyperparameters Tuning

Regularization is the method used to **reduce the overfitting** of training data.

`sklearn.linear_model.LinearRegression`

```
class sklearn.linear_model.LinearRegression(*, fit_intercept=True, copy_X=True, n_jobs=None, positive=False)
```

[\[source\]](#)

```
from sklearn.linear_model import LinearRegression
reg = LinearRegression().fit(X_train, y_train)
reg.predict(X_test)
```

- Lasso for L1 regularization
- Ridge for L2 regularization
- Elastic Net regression for both L1 and L2

```
from sklearn.linear_model import Lasso
```

```
lasso_reg = Lasso(alpha=0.1)
lasso_reg.fit(X, y)
```

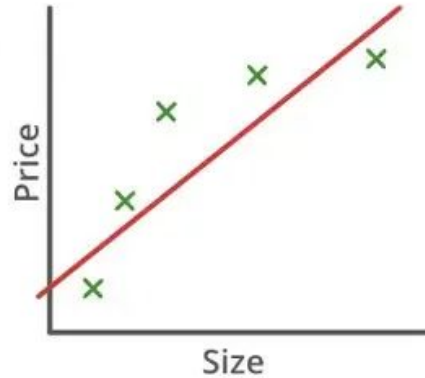
```
from sklearn.linear_model import Ridge
```

```
ridge_reg = Ridge(alpha=0.1)
ridge_reg.fit(X, y)
```

```
from sklearn.linear_model import ElasticNet
```

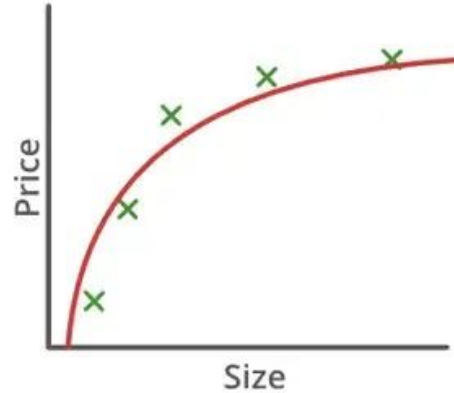
```
elasticnet_reg = ElasticNet(alpha=0.1, l1_ratio=0.5)
elasticnet_reg.fit(X, y)
```

ML | Underfitting and Overfitting



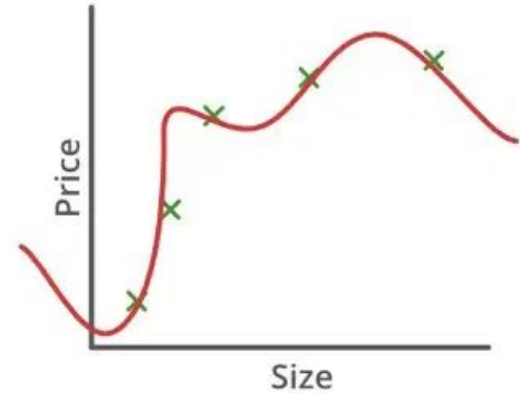
$$\theta_0 + \theta_1 x$$

High Bias
(Underfitting)



$$\theta_0 + \theta_1 x + \theta_2 x^2$$

Low Bias, Low Variance
(Goodfitting)



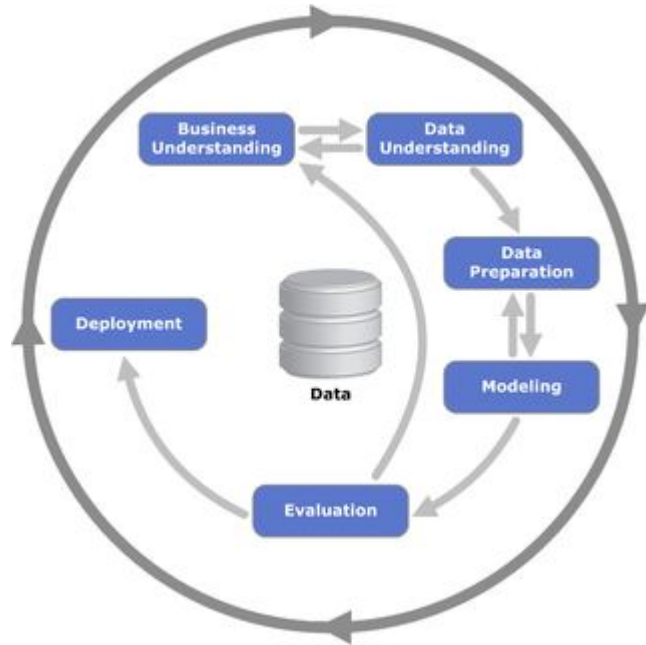
$$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

High Variance
(Overfitting)



8. Explain the model and Business Constraints Using CRISP-DM

[Link to Experiment Report](#)



1. EXPERIMENT BACKGROUND	
Provide information about the problem/project such as the scope, the overall objective, expectations. Lay down the goal of this experiment and what are the insights, answers you want to gain or level of performance you are expecting to reach.	
1.a. Business Objective	Explain clearly what is the goal of this project for the business. How will the results be used? What will be the impact of accurate or incorrect results?
1.b. Hypothesis	Present the hypothesis you want to test, the question you want to answer or the insight you are seeking. Explain the reasons why you think it is worthwhile considering it,
1.c. Experiment Objective	Detail what will be the expected outcome of the experiment. If possible, estimate the goal you are expecting. List the possible scenarios resulting from this experiment.

Machine Learning Exercise

Please work in a team of 3-5 people to address a Linear Regression problem of predicting house prices using “USA_Housing.csv” dataset. By following these steps:

1. Spend the first 10 minutes discussing
 - a. Business Objective,
 - b. Hypothesis,
 - c. Experiment Objective
2. Explore and Preprocess Data
3. Train a Linear Regression, Lasso, Ridge, and ElasticNet
4. Choose which model to use and discuss why

https://github.com/merrymira/UPASS_ML_WEEK3



Solution

Link to Github

https://github.com/merrymira/UPASS_ML_WEEK3

