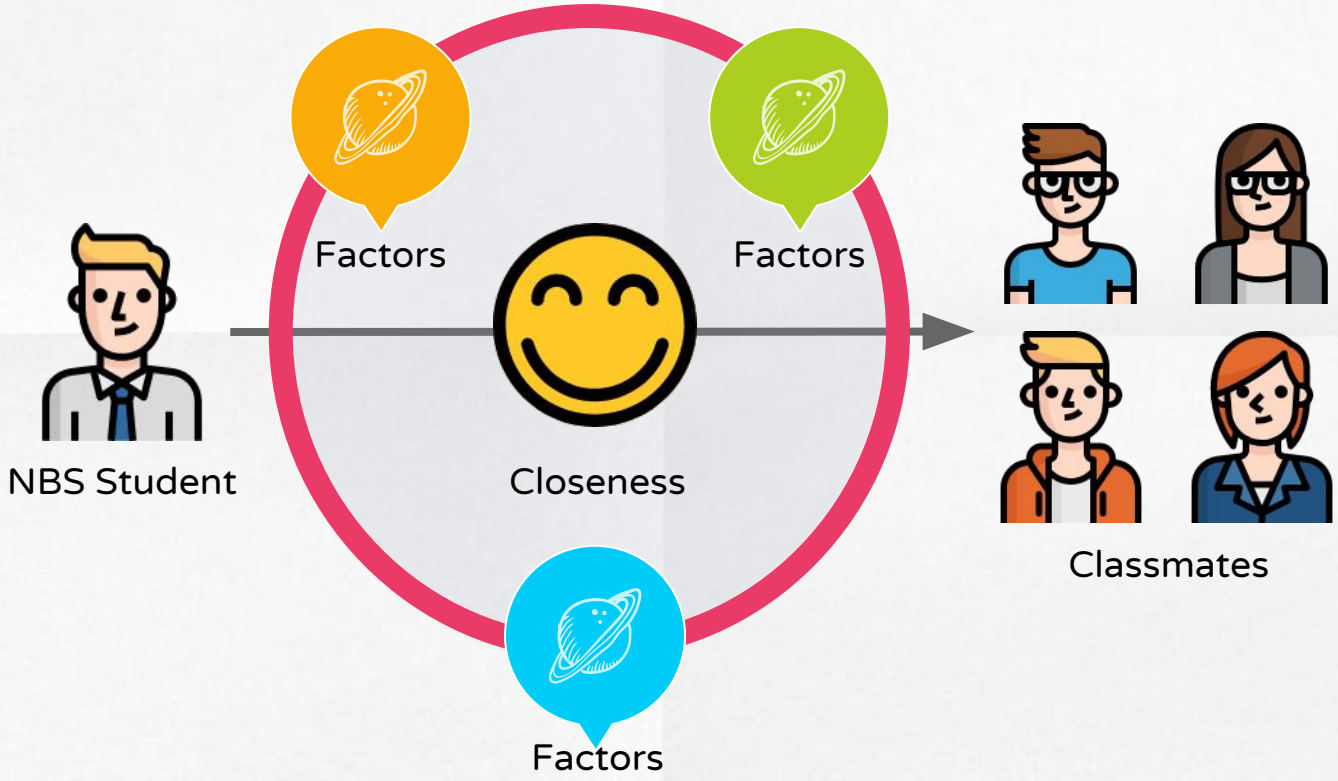
How Close Are NBS Students To Their Classmates?

Cheng Jin Yee (**Jinny**)
Jeremy Jerome Chia

CE9010 Introduction to Data Science
AY1718 S2

1. Data Problem



2. Data Acquisition

Survey

"Friends in NBS"

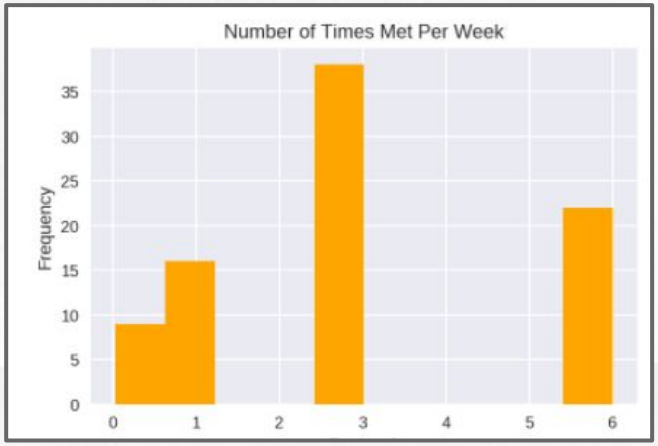
Factors

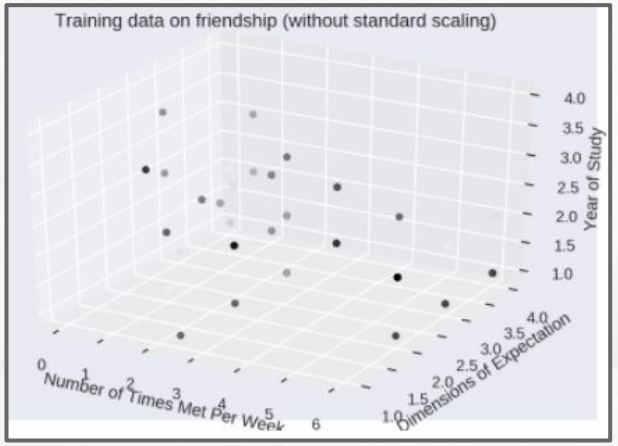
How Often Do You Meet?

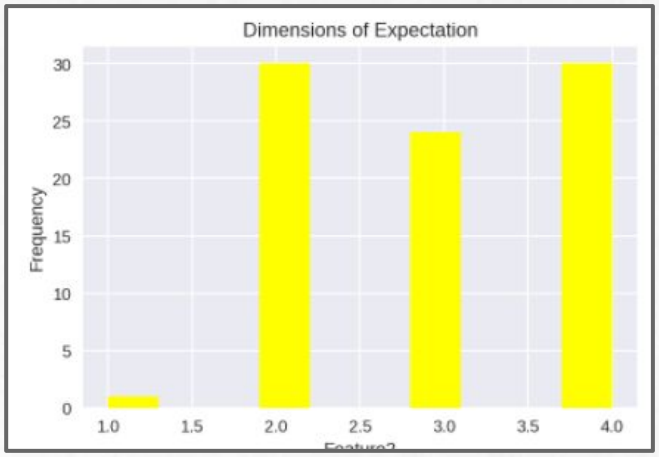
Dimension of Expectations

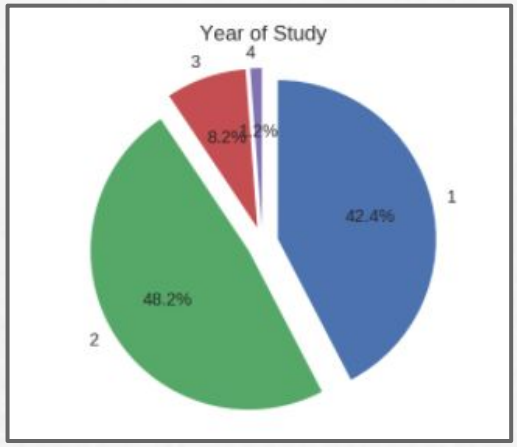
Year of Study

3. Data Exploration

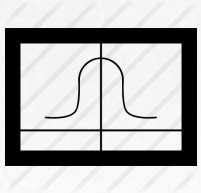






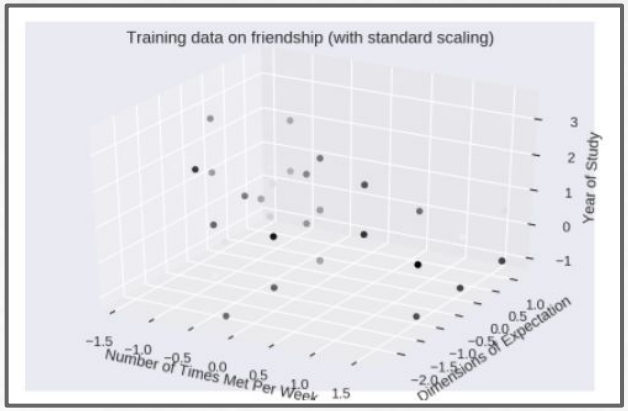


4. Pre-Processing



z-Scoring of Data

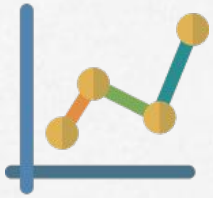
$$z_i = \frac{(x_i - \bar{x}_i)}{\sigma_i}$$



Removal of Noise Data

What makes me feel close to my frie	How often do I s	Le	ssm	cCan
Common interest/topic	Every week		6	6
.	2-3 Times a Month		8	3
same timetable	2-3 Times a Week		8	5
Our friendship	Every week	kkwu	10	8
their sincerity	2-3 Times a Week		10	5
Things done	2-3 Times a Week		3	1
It makes me feel close to my friends when	Almost every day	yloke	10	9
More time together	Almost every day		9	7
Frequent interaction	2-3 Times a Week	ctay0	10	2
They cares about me	Almost every day		6	1
Time spent hanging out, common talking t	2-3 Times a Week		3	0
depth	2-3 Times a Month		7	2

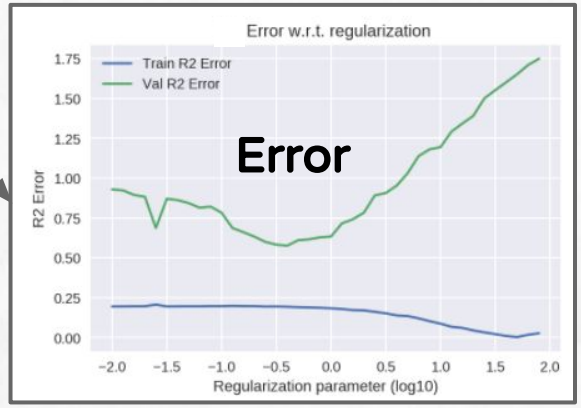
5. Data Analysis



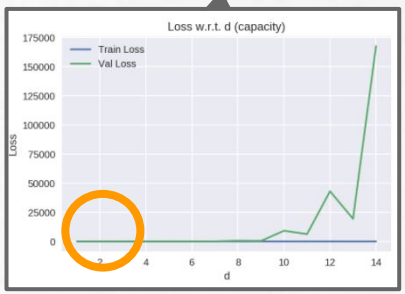
Regression Analysis

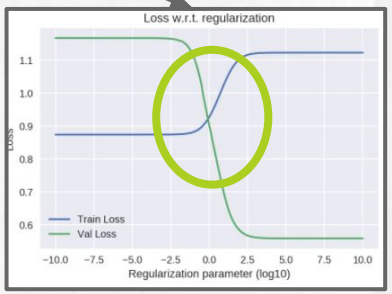
$$y_i = w_1 x_1 + w_2 x_2 + w_3 x_3 + \ldots$$

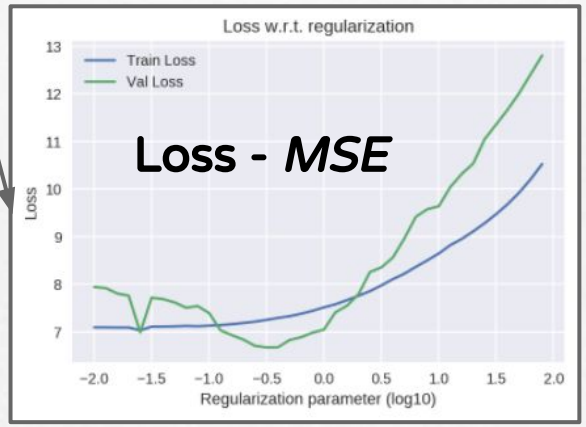
Evaluated By



Techniques Used



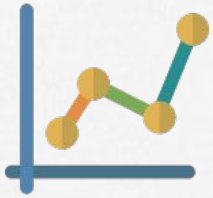




Overfit training data and finding balance

Add regularisation to the validation set and evaluate generalisation of the validation set

5. Data Analysis



Regression Analysis

$$y_i = w_1 x_1 + w_2 x_2 + w_3 x_3 + \ldots$$

- Without Normalisation
- Reducing Numbers of Features
- K-Fold Cross Validation

Techinque	Optimal d	λ	Train Loss	Val Loss
With Normalisation	2	0.794	0.91	0.92
Pre-processing	2	0.316	7.25	6.67
Removal of Feature 1	3	0.398	1.00	1.01
Removal of Feature 2	2	0.316	0.94	0.94
Removal of Feature 3	2	1.000	0.98	0.68
K-Fold Cross Validation	1	0.100	0.87	0.85

BEST MODEL!

6. Analysis of Results

Observations

- **Normalisation** increases the performance of the model.
- **Regularisation** strikes a balance between bias and variance, and generalises the well-fitted model.
- **Cross-Validation** reduces the loss.
- **Data Features** may not be the most suitable, though there is evidence of some relationship.
    - Loss does not go below 0.6 for the training data even after overfitting; may not be the most appropriate features.

Improvements

- Collect **More Data**
    - Reach maximum learning capacity
    - At high capacity, collecting a large number of data will not cause overfitting and can greatly reduce loss (fix high variance)

- Select more **representative features**
    - Handcrafting Features
    - Fix High Variance


The END