CS559 – Project 1
Due: 3/10th Wednesday 11:59 PM

Project: Clustering & Dimension Reduction
Description: With a given data, the data provider wants to reduce the size of data (dimensions) before the modeling. Under the assumption that the provider does not know the details except that it has 200 features yet, you have the freedom of handling data – there are no specific ways of reducing the dimensions.

Data Description:

- Size – 201 columns and 200,000 rows.
- Columns – Except the first column "ID_code", the rest columns are not labeled.
- Target – unknown.

Task Ideas:

1. EDA – Considering the data size, it will not be a great idea to cluster or reduce dimensions using the original data. Therefore, by doing EDA, find out if there are any unique characteristics among columns and between columns.
1. Feature Engineering & Extraction – Using learning from EDA, determine columns can be used to create features or can be eliminated for further analysis.
2. Clustering Analysis – Determine the appropriate cluster number for this dataset. Considering the number of observations, you have to be creative on methodology to have the lower computational cost.
3. Dimension Reduction – Using the principal component analysis, determine how many principal components can be taken.

Submission:

- Code – can be submit as HW – HTML and ipynb.
- Report – 4-5 pages formal report is required. The contents of reports are
    - Introduction: Do not need to explain the background of techniques or functions used unless if an author feels strongly. Have a brief explanation of work and the paper.
    - Methodology: Explain how the tasks were attacked and reasons the importance of steps.
    - Results: Explain what you found from the above section
    - Discussion & Conclusion: Evaluate the work, suggest alternative ways can be done, and report the final dimension the author like to propose to the data provider.

Grading:

- Work effort (40%)
- Thorough explanations of work in Jupyter Notebook (20%)
- Report (30%)
- Quality of result – is it convincible? (10%)