# Peer Review For Group 7

Mini-Project 1. MATH 4995

Group Number: 7

Title: Supervised Classification with Full Spectrum or Premium Subset?

## Summary of the report

The report aims to predict if a client will default in Home Credit Default Risk problem. Moreover, the report will investigate if only a partial features is selected, would the performance of the model better than little to no features is being eliminated. Features are removed based on multicollinearity. From the investigation, it has been found that partial features would perform better in logistic regression, while the performance would not differ much if LDA and random forest is used.

## Strengths of the report

The report proposed a question and applied it through different model, which gives out different result. Therefore, instead of giving one general result, the result given would be more diverse and meaningful. Moreover, a strategic way of choosing features makes the reasoning of choosing specific subset better. Some of the data analysis have made them aware of some data with non-common distributions, which could potentially improve the feature selection.

## Weaknesses of the report

As only the training and testing table is used, and that other tables are discarded, it might be possible that some of the features aggregated might be more useful than the features in the training or testing table. Moreover, it might be necessary why it is 5-fold cross validation, instead of cross validation of different values. Moreover, since some of the results is already known (curse of dimensionality affect some models more than others), it might be better to indicate some past research result on that matter.

## Evaluation on quality of writing: 5

Is the report clearly written? Is there a good use of examples and figures? Is it well organized? Are there problems with style and grammar? Are there issues with typos, formatting, references, etc.? Please make suggestions to improve the clarity of the pap er, and provide details of typos.

The report is very informative. Each of the section are in details. The goal of the project (if subset of features are useful) is clearly written, and that the whole report follows this question. Feature selection would be an important part of this project rather than model selection, which in this report the feature selection part is in detail. There are a variety of charts, including distribution charts, explained variance, which could helps feature selection or pre-processing from different perspective. However, I believe that the selection of why those 16 features are selected could be written with more detail, for example, is it only based on correlations, and are there other factors other than correlation (For example issues on feature correlation with other features).

## Evaluation on presentation: 4

Is the presentation clear and well organized? Are the language flow fluent and persuasive? Are the slides clear and well elaborated? Please make suggestions to improve the presentation.

The presentation is quite well organized. The presentation follow the logical flow which makes it easy to follow. Many of the visuals, including graphs on certain data, and some images used helps us to understand the presentation. The analysis is also detailed enough on result that is not only applicable to this specific data, but seems to be generalizable also, which could be useful as it is possible to do follow-up investigations. It would be better if there is more comparison between different model, in terms of explaining why some models may perform differently with different amount of variables fed into it while some models performs the same.

## Evaluation on creativity: 3

Does the work propose any genuinely new ideas? Is this a work that you are eager to read and cite? Does it contain some state-of-the-art results? As a reviewer you should try to assess whether the ideas are truly new and creative. Novel combinations, adaptations or extensions of existing ideas are also valuable.

The report did give out inspiration on whether or not feature selection is necessary for some models. For example, doing feature selection in logistic regression can help improve performance, and it is not necessary to do feature selection for LDA and random forest. While quite amount of people in the Home Credit Default Risk problem did do feature selection (as generally it performs better) this result did give us information on its usefulness.

## Confidence on your assessment: 3

(3- I have carefully read the paper and checked the results, 2- I just browse the paper without checking the details, 1- My assessment can be wrong)

The poster, presentation and source code are clear enough to understand the entire flow of the work, which also makes me confident on the assessment.