Rebuttal from group 4

Review 4:
Rebuttal on merging dataset:
We did mention the structure of datasets in P.6 of our presentation and Part 2 of our poster. Detailed implementation can be referred to our code, as most of the codes are basic data manipulation techniques.

Rebuttal on feature selection:
First, the dataset provider did not give a clear explanation for every attribute, e.g. the three most important features EXT_SOURCE_1&2&3, are just described as Normalized scores from external data sources.
Second, we did examine the significance of each feature base. Although the dataset provider did not give a clear explanation, we have checked from the discussion that EXT_SOURCE_1&2&3 are actually credit ratings from several credit rating agencies. And we have explained it in our presentation.
Third, we have done data exploratory analysis on the relationship between features and response variables. It is also clearly explained and included in part 9 - 14 in our presentation slides.
Fourth, as for the pairwise scatter plot, we did try to conduct it, however, due to the large number of attributes, it was unfortunate that we were unable to conduct the pairwise scatter plot for each attribute.

Rebuttal on quality of writing:
We presented clearly using figures and graphs, with the ROC curve as a comparison of accuracy of logistic regression versus lightGBM. Moreover, we tried both application dataset and merged dataset for Logistic Regression and LightGBM. It is clearly identified in P.22 of our presentation slide and part 5 of our poster.

Rebuttal on creativity:
We proposed a novel problem statement about analyzing the difference between merged dataset and application dataset, the focus of this project is to examine whether more features can give higher prediction results. Therefore, we adopted two major models to test this hypothesis. And Light GBM is a state-of-art model which was just developed in recent years. It has a high model complexity and accuracy, and it is the most winning model in Kaggle contests. Moreover, we did spend great effort on understanding the deep architecture, explaining to the class instead of just using it. So we think that these two major models are sufficient to draw our conclusion that more features are indeed useful for prediction. Moreover, we did tune our LightGBM model, with fine tuning parameters and grid search for hyperparameters, as presented and explained in P.20 of our presentation slides. Detailed implementation can be referred to our code. We do not agree that we just use it in a very simple way. We believe we should deserve a higher score in this part.