1. Summary of the report.
The group tries to find out the best predictors according to the correlation
coefficient calculation and use logistic regression to train the algorithm. Kaggle
test result gets 0.69290 private score and 0.70703 public score.

2. Describe the strengths of the report.

Labeling:
The group uses one hot encoding to label variables that has more than 2 categories.
If the column has no more than 2 unique categories, it is labeled 0 and 1.

Data Cleaning:
There are columns in the training set but not in the test set (besides Target). Only
those appear in both sets are kept and the rest are dropped. Missing values are
replaced by medians.

Variable Selection:
The most significant predictors are selected by comparing the correlations. The
method is reasonable since the predictors affecting the response variable most
should have a significant correlation with it as well. The advantage of the method
is that it is easy to implement. One thing that surprises me is that among hundreds
of predictors, finally only 4 are selected to perform the logistic regression. The team
has its own thoughts about which variables should be selected, rather than blindly
taking all of them into the regression.

3. Describe the weaknesses of the report.

I think the biggest weakness of the method is the lack of interpretability. They use
cubic polynomial features to perform the logistic regression. It improves the ROC
result of the logistic regression by a great amount, compared with the

one with
the original predictors. However, it becomes hard for us to interpret why and how
those transformed cubic polynomial features should have predictive power over
the response Target. In the report this point is mentioned in the conclusion, that
they cannot get a well explanation of their algorithm, because of the indistinct
data description of some of the features they used.

The ROC curve (on the training set) with the original 4 predictors performance is
only slightly better than 50/50. I think perhaps they dropped too many variables
in the first place. Besides the 4 predictors, there may be more significant variables
that should be included in the logistic regression.

4. Evaluation on Clarity and quality of writing (1-5): 5

5. Evaluation on Technical Quality (1-5): 4

The introduction of cubic polynomial feature demonstrates novelty. Overall, the
experimental results are well-supported.

The final score on Kaggle is around 0.7, which I think is around average and there
is room for improvements.

6. Overall rating: 4- A good report.

7. Confidence on your assessment (1-3)
3- I have carefully read the paper and checked the results