**MSBD 5013 Statistical Prediction**

**Project 1 Rebuttal**

CHENG, Pak Hei(20795962)

LO, Ngai Hung(20787771)

**Selected comments for rebuttal:**

* *“Although they discussed the issue of imbalanced label, they did not take actions to solve it thoroughly. One possible improvement might be oversampling and desampling” - Pengzhixuan*  
    
  Author’s reply:  
  To handle imbalanced data, we removed heavily imbalanced data and tried to apply XGBoost and LightGBM, which are generally considered as algorithms that can provide good performance on imbalanced data. Resampling methods are good ways to handle imbalanced data, we used our models’ parameters such as “scale\_pos\_weight” to introduce some sort of importance to the minority class in the loss function. This way we may avoid overfitting by oversampling and make use of all the data. Implementation details can be found at “model\_search.ipynb”.
* *“Report written that used coefficient to subtract the feature dimensions, it seems not in the code” - Yangyuxin*  
    
  Author’s reply:  
  In the “Data\_pre\_and\_model.ipynb” file, Pearson correlation analysis was done and features with low correlation with “Target” are manually added into a list called “drop\_column” and removed later.
* *“It seems not to talk about the numerical feature and categorical feature separately” - Yangyuxin*  
    
  Author’s reply:  
  We handled both data types in a single data cleaning and engineering flow. Therefore, we mentioned the key strategy adopted in this flow rather than discussing how we handle numerical features and categorical features separately. We focused more on data analysis and model interpretation in the report and expected the readers to discover engineering details in our code.
* *“Besides, during the “pre-processing” part in the code, it seems there is no reason for creating some feature” - Yangyuxin*  
    
  Author’s reply:  
  We hope to get more specific feedback from the reviewer on which features were unnecessary and why they were. As over 300 features were used, we would like to understand how the reviewer deduced some features were unnecessary. We do agree that some features will result in no impact to model performance, which is why we used correlation analysis to eliminate those features.
* *“They didn’t explain why there are some different performances between these tree models. In other word, if they assessed the strength and weakness of their models, the work would be more completed.” - Wuxiang*  
    
  Author’s reply:  
  It is true model comparison analysis is lacking in our report, we appreciate the comment by the reviewer. It is also in our interest to conduct such an analysis in the future.