**The level of customer loyalty**

ln this program, data is from Professor Jagdip Singh at Weatherhead School of Management from Case Western Reserve University. Programming by Jiaqi Li, Zeyao Yang and Ziming Cui.

***1. Usage Instructions***

If you wish to get a prediction result, change the dictionary paths for the data files containing the training and test data. Next, modify the output dictionary to point to the desired place. The software will output the predicted outcome.

***2. Background***

The level of customer loyalty will greatly affect the profitability of Bancare, so this program intends to collect customer data on LOYALTY and build a predictive model using a range of predictors as described in the “data-description” file.

***3. Methodology***

*To develop a predictive model that:*

1. Has high Accuracy, Recall, Precision and F1-score 2. Few predictors 3. Is robust and not biased by violations 4. Includes important interaction and quadratic terms.

Firstly, divide the data at random into a training set (used to develop a predictive model) and a test set (used to evaluate the model). Calculate penalized logistic regression with Lasso Regression to automatically choose the best model with the most useful predictor variable.

And then, utilizing cross-validation to select the best lambda to do feature selection. To choose the crucial predictor variables and interaction terms, run the ANOVA test. Using normalized predictor variables, construct an LDA model.

Compare the accuracy achieved with Lasso Regression to that achieved with the whole logistic regression model (containing all predictors) and LDA model. Analyze significant performance indicators, including classification-model assessment and confusion metrics (ROC and AUC).

Finally, we added important interaction (Value of use × reputation) and quadratic terms, comparing LDA model, logistic regression with a reduced set of variables. Therefore, we choose the LDA mode as the final best model, and the final accuracy is 0.827.

***4. Other Concerns***

Our model still exists multicollinearity between independent variables, which results in less reliable statistical inferences, therefore limit the research conclusions we can draw.

***5. Conclusion***

Logit(Y) = 0.2948 + 0.8772 \*Satis + 1.0504 \* Value + 1.6845 \* Reputa – 1.2017 \* sex\_male + 0.4751 \* CTRUST – 0.9256 \* (Reputa \* sex\_male) – 0.8668 \* (VALUE \* Reputa)

The final best model has high Accuracy, because of its accuracy (0.827), Recall (0.829), Precision (0.866) and F1-score (0.847). And we delete most low impact predictors like age1, age2, etc. which play an important role in making our model robust and not biased by violations, including important interaction (Reputa: sex\_male and VALUE: Reputa).