Northwestern | McCORMICK SCHOOL OF ENGINEERING

MSIT 423

Data Mining & Business Intelligence

June 9, 2018

Book-Case (Customer Spend Model)

Kaggle Group/Team Name: **MSIT_United Nations**

Group Members**:**

- Ruma Anand
- Akshay Jain
- Ademola Popoola
- Guanglong Xu
- Liping Zhou

Best Predictor Variables Across Models :

- Time_per order - tott/ford [#tott = total time spend]
- Money spent per order - m/ford
- tof + r + ford + f1 + f5 + f6 + f9 + f14 + f17 + f39 + m + f8 + f19 + f5:m + tof:f8 + m:f19

Best Working Models :

Our best working model was a Two -Step + Stepwise model which is mentioned below:

- fit=glm(formula = logtarg > 0 ~ tof + r + ford + f1 + f5 + f6 + f9 + f14 + f17 + f39 + m + f8 + f19 + f5:m + tof:f8 + m:f19, family = binomial, data = all[train, 3:38])
- fit2=lm(formula = logtarg ~ r + sqrt(r) + m + sqrt(m) + ford + f5 + f8 + f19 + f36 + f19:tof, data = all[train,], subset = (logtarg > 0))
- yhat=predict(fit,all[!train,],type="resp")*predict(fit2,all[!train,3:38])
- sqrt(mean((all$logtarg[train] - yhat)^2))

“Not So Good” Models:

- fit = lm(logtarg ~ ., all[train, -c(1,2)])
- No transformation + Ridge/Lasso
- No transformation
- Model with m and fitem (both were collinear)
- Model without RFM variables (m, ford, fitem, tof, r)

“Secrets” to Group’s Success:

- Two Step model + Stepwise