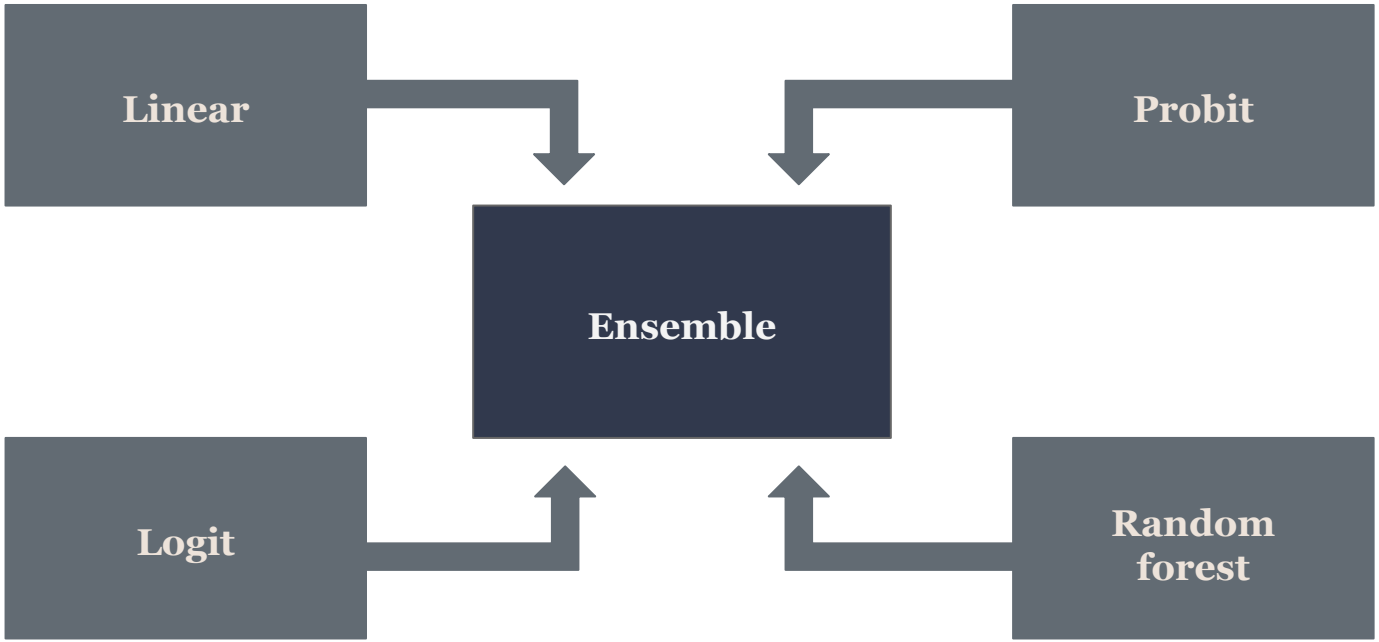
ETC3250 Business Analytics
Project Presentation

ggplotters2

Yuanyuan Chang
Alex Do
Chairach Kraisarin
Yan Li
Christopher Trinh
Yangzhuoran Yang

Ensemble model



Model formulation

Linear regression/LPM

1. Best Subset Selection
2. Choose model with maximum Adjusted R^2 and minimum Cp

Logit and probit

1. Mixed stepwise variable selection (start at full model then add or drop variables)
2. Choose model with minimum AIC

Random forest

1. Automatic variable selection
2. Choose parameters (number of trees, number of node predictors) - use validation set approach (20/80 split) with log loss function

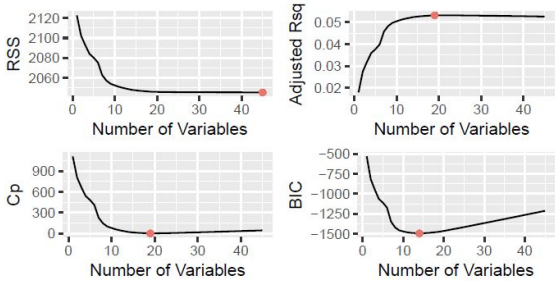


Figure 1: Plot of RSS, Adjusted R^2, C_p and BIC against number of regressors

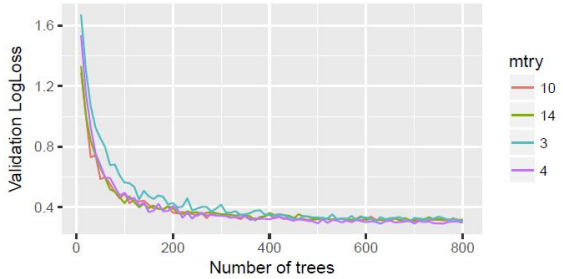


Figure 2: Plot of validation set log losses against number of trees, with number of node predictors.

Results and discussion

1) Improvement of log loss using ensemble modelling

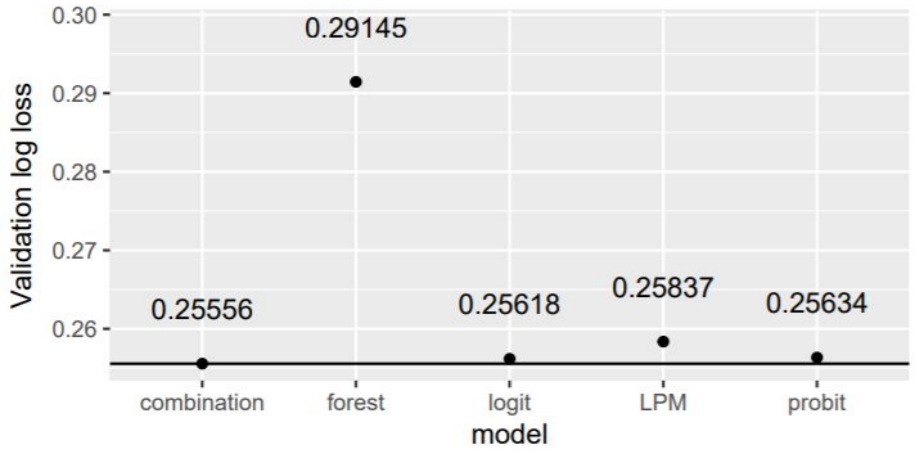


Figure 3: Validation log loss for different models

Results and discussion

2) Effects of variables on different models

Using **t-test** on LPM and Logit & Probit Model

❖ Linear Regression Model

Important variables:

- Employment Status
- Date of last contact
- Number of contact during campaign
- Credit in default

❖ Logit and Probit Model

Important variables:

- Employment Status
- Date of last contact
- Number of contact during campaign
- Number of day from previous campaign
- Out come from previous marketing campaign

Results and discussion

2) Effects of variables on different models

- **Random Forest**

Gini Index: $$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

Important variables:

- Age
- Employment status
- Date of the last contact
- Number of contact during campaign
- Education level

Table 1: Variable importance in random forest

	MeanDecreaseGini
age	740.26862
job	374.64890
marital	159.65267
default	81.74803
housing	156.38224
loan	117.60036
contact	39.34858
month	326.24650
day_of_week	306.45773
campaign	376.89208
pdays	17.82658
previous	27.70053
poutcome	29.67767
edu	309.75355

Limitations/next steps

- Limited computing power prevented the use of k-fold cross-validation with high values of k when selecting random forest parameters.
- Using different weights when ensemble averaging to reduce/offset the high error from random forest.
- Exploring boosting instead of random forest
- Exploring the use of shrinkage methods such as ridge/lasso regression

Thank you!

Any questions?