Stat 412

Yolanda Jin

24/11/2021

EDA

- Might need to filter out NA data
- monthly income = 0
- 30k monthly income = NA
-

Dependent = NA 4k

- age 0 remove - only 1 point
- age very old
- group by age group etc
- 13 - 101 or older (maybe cut at 100)
- 80 or more

Question

Comparing our model against paper's model

- Most important factors

Ensemble Learning

- Lasso Ensemble Algorithm
- Aggregating base learner: Weighted base=learner

Balancing Data

- Use clustering and pick one sub-group from majority
- Use bagging algorithm to create more minority data
- Use NA indicator 0 or 1 (maybe use mean/median)

Which models to try

- Logistic regression (compare the different link functions)
- look maybe merge Lasso with logistic regression
- RF

Evaluating Model/Comparing results

- AUC