Kaggle: BNP

Dat Tran, Gerrit Gruben

February 6, 2016

- Binary classification problem. Incoming claim needs human intervention y/n?

- Binary classification problem. Incoming claim needs human intervention y/n?
- 131 (21 categorical, 110 real) features, $115k$ training data points.

- Many missing values, average non-NaN density is around 0.66.

- Many missing values, average non-NaN density is around 0.66.
- Categorical variables have many values $\Rightarrow$ One-Hot-Encoding will need sparse data structures.

- Logistic Regression not viable, because of missing values.

- Logistic Regression not viable, because of missing values.
- Used Tree-models (Random Forest) to deal with it, dropped categorical features.

- Logistic Regression not viable, because of missing values.
- Used Tree-models (Random Forest) to deal with it, dropped categorical features.
- Split 30% away for seperate holdout. RF without any tuning performs with log-loss of about 9.13 and accuracy of 73.5%.

- Use grid search to find params, estimate log-loss from CV.

- Use grid search to find params, estimate log-loss from CV.
- This RF is expected to be a first good submission (keep it for ensembling later)

- Use grid search to find params, estimate log-loss from CV.
- This RF is expected to be a first good submission (keep it for ensembling later)
- Use categorical variables, analyse feature importance.