Titanic: Machine Learning from Disaster

Goal: To predict what kind of people were more likely to survive

Variable Description:

VARIABLE DESCRIPTIONS:

survival Survival

(0 = No; 1 = Yes)

pclass Passenger Class

(1 = 1st; 2 = 2nd; 3 = 3rd)

name Name

sex Sex

age Age

sibsp Number of Siblings/Spouses Aboard

parch Number of Parents/Children Aboard

ticket Ticket Number

fare Passenger Fare

cabin Cabin

embarked Port of Embarkation

(C = Cherbourg; Q = Queenstown; S = Southampton)

Raw predictors: pclass, sex, age, sibsp, parch, fare, cabin, embarked

Synthetic variables:

* Family size = sibsp + parch + 1
* Title. Extract title information from name (Miss, Mrs, Mr, Master)
* Cabin. Classify missing value to a new category

Interaction variables:

* age\*sex
* age\*fare
* pclass\*sex
* pclass\*age
* etc.

Phase 1: Simple Modeling (1/24 – 1/25)

- Load Training and Testing datasets and combine the two datasets

- Check the structure, classes, distribution, outliers

- Impute the missing values

- Create synthetic variables

- Transform factors to numeric indicators

- Check the correlation between target variable and predictors

- Simple modeling (GLM, CART)

Phase 2: Advanced Modeling (1/26 – 1/27)

- preprocess: check the correlations between predictors

- Model building (GLM, CART, Random Forest, GBM)

- Evaluate model performance using lift, confusion matrix, accuracy, AUC, etc.

Phase 3: Model Refining (1/30 – 2/1)

- Read and learn from the examples

- Refine the modeling work and finalize the results

Phase 4: Documentation (2/2 – 2/4)

- Use the best model for documentation purpose