Model building on Bank Marketing Campaign

Shuran Fu Aug 18, 2022

Agenda

Problem Statement

Data transformation

Model Building and Selection

Recommendations

Summary

Problem Statement

- Develop a machine learning model to estimate whether a particular customer will buy a specific term deposit product or not based on the customer's past interaction with bank or other Financial institution.
- y in dataset means whether the customer buy the product or not, so this is a supervised learning problem and we need to use classification model.
- data cleaning -> EDA -> Feature selection -> Model construction -> Performance analysis

Data Transformation

- change binary categorical variables into 1 and 0
- change multi-categorical variables into numbers
- change string data type into integer
- split dataset into train set and test set with fixed random state

Model Building and Selection

- Base Model - Logistic Regression
- Random Forest
- LDA
- Decision Tree
- Boosting methods
- Neural Network

Model Building and Selection

model	in-sample error	out-of-sample error
logistic regression	7.49	7.44
random forest	6.24	6.41
lda	7.51	7.45
decision tree	6.29	6.38
adaboost	6.77	6.80
gradient boost	5.87	6.40
neural network	7.45	7.5

Model Building and Selection

- Use pearson coefficient to determine features
- feature group 1: job, education, balance, housing, loan, duration, campaign, previous, poutcome
- feature group 2: housing, duration, campaign, previous, poutcome

Model Building and Selection

model-feature group 1	in-sample error	out-of-sample error
random forest	6.18	6.31
decision tree	6.34	6.25
gradient boost	6.04	6.41
model-feature group 2	**in-sample error**	**out-of-sample error**
random forest	6.14	6.38
decision tree	6.34	6.25
gradient boost	6.15	6.35

Recommendations

- Use decision tree or gradient boost classifier to predict whether a particular customer will buy a specific term deposit product or not
- Adding or deleting some less important features do not generate significant impact on the result of the model
- Confusion matrix is also important if we want to know the false positive rate and false negative rate

Summary

- Problem statement
- Data preprocessing
- Model building and metrics evaluation
- Recommendations

Thanks !