A complete analysis for predicting banking behavior by the classification method.

By Huanwang (Henry) Yang (2021-6-5)

Objective:

Build a predictive model to score each potential customer's propensity of subscribing a term deposit, as well as understanding which customer characteristics are most important in driving purchasing behavior, in order to inform future marketing segmentation personalization.

Dataset:

The data was obtained from the UCI data resource http://archive.ics.uci.edu/ml/datasets/Bank+Marketing

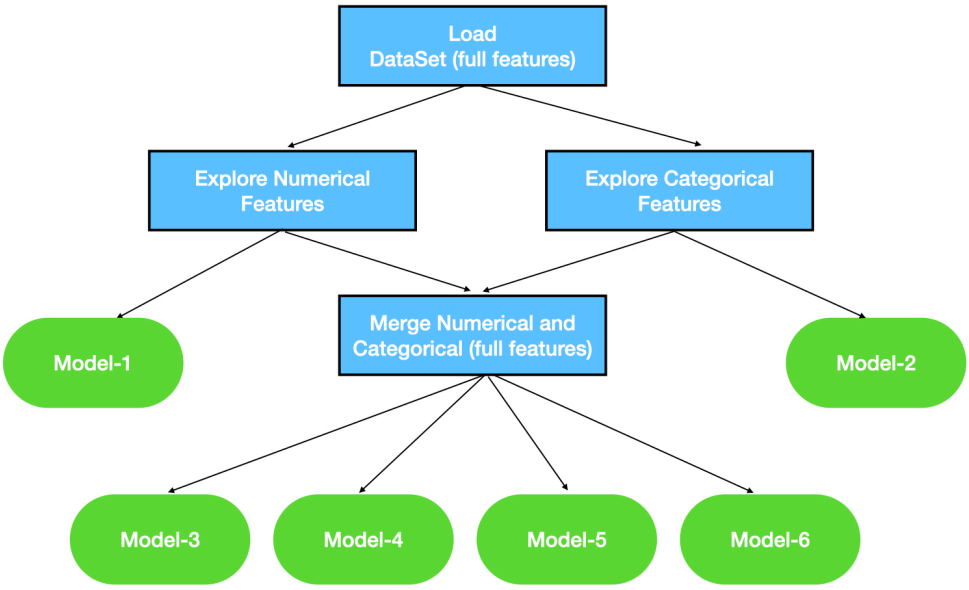
Modeling procedures:

The XGBoost, CNN (CONV1D) and LogisticRegression models were built for binary classification analysis. I used different datasets to test the differences among the models.

Note: the attribute 'duration' has been removed from all the models due to the reason as described in the original datasets:

last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

The diagram below shows the modeling procedures in the jupyter notebook



Model-1: XGBoost using data for numerical features only
Model-2: XGBoost using data for categorical features only
Model-3: XGBoost using data for the full features (numerical + categorical)
Model-4: XGBoost using data for the full features, but balancing the two classies using SMOTE.
Model-5: CNN (CONV1D) using data for the full features
Model-6: LogisticRegression using data for full features (use statsmodels for details coefficients)

A summary of result:

Model:	AUC_ROC	Accuracy	precision	recall	f1_score_mean
Model-1	0.87	0.80	0.70	0.74	0.72
Model-2	0.88	0.77	0.70	0.68	0.69
Model-3	0.87	0.80	0.69	0.75	0.72
Model-4	0.86	0.79	0.68	0.74	0.70
Model-5	0.90	0.79	0.80	0.58	0.61

From all the modelings, it looks like the model-3 is among the best. It is interesting to note that model-1 and model-2 do not differ a lot.

Feature importance

The three features (below) are the most important factors for driving purchasing behavior.

feature	importance
nr.employed	0.343
poutcome_success	0.130
emp.var.rate	0.103

Conclusions:

This classification can help the manager to make business dicision based on the modeling result. The feature importance shows which customer characteristics are most important in driving purchasing behavior. The coefficients (along with the P-values) from LogisticRegression (model-6) provide a quantitive probability for the customer characteristics.