June 2024

Offer Acceptance Prediction

Bartosz Grabek, Izabela Telejko, Grzegorz Zbrzeżny

Agenda

- Project Objectives
- Methodology
- Selected Experiments
- Results and Final Strategy

Project Objectives

Project Objectives

- develop a **strategy** (feature selector + binary classifier) that classifies the customers that would benefit from a bank marketing offer

- use the developed model trained on historical data to find the 1000 customers out of 5000 customer test data most likely to accept a bank marketing offer

Methodology

Evaluation / Scoring

$$score = 10P \cdot \frac{|X|}{|X_{val}|} - 200N,$$

where:

- $|X|$ – size of the whole dataset,
- $|X_{val}|$ – size of each validation dataset in cross-validation,
- P – number of properly classified clients in validation dataset in given split out of $\frac{|X_{val}|}{5}$ clients with highest probability of benefiting from the offer (division by 5 comes from the fact that in our main task we select 1000 clients out of 5000),
- N – number of features selected in given split.

Selected Experiments

Feature Selection Methods

- CMIM – Conditional Mutual Information Maximization
- JMIM – Joint Mutual Information Maximization
- IGFS – Information Gain Feature Selection
- Random Forest Feature Importance (Impurity Decrease and Permutation Importance)
- Boruta
- Chi-Square Test
- Fisher score

Classification Models

- Support Vector Machines (SVMs)
- Logistic Regression
- LDA and QDA
- Naive Bayes
- Tree-based methods such as Random Forest, AdaBoost, Gradient Boosting
- Multi-layer Perceptron
- Ensembles and stacking of the best models

Results

Selected strategies

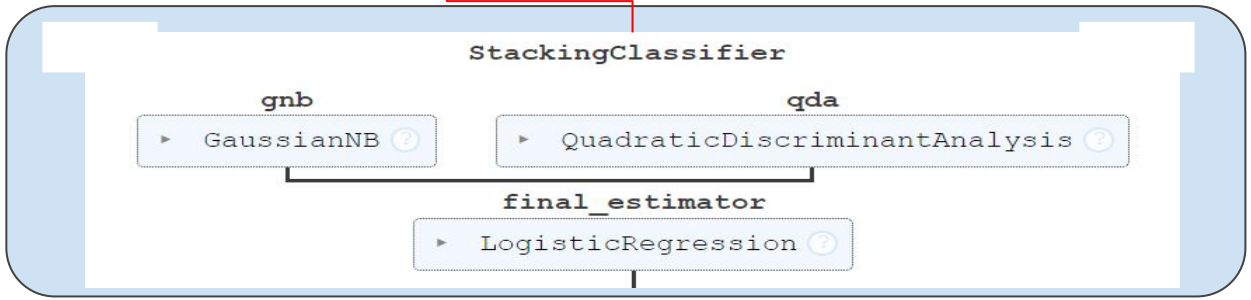
Feat. Sel.	# Feat.	Model	Average Score
CMIM	4	QDA	5810
	3	SVC	5580
	3	MLP	5490
Boruta + JMIM	4	QDA	6230
	3	SVC	6040
	4	MLP	6010
Boruta + IGFS	3	QDA	6270
	4	SVC	6170
	4	MLP	6130
Impurity	4	GaussianNB	6880
	4	QDA	6850
	4	MLP	6850
Boruta + Permutation	3	GaussianNB	6840
	3	SVC	6800
	3	QDA	6830

Commonly Selected Features

Feat. Sel.	commonly selected features
CMIM	1, 2, 3, 5, 6, 22, 28, 30, 39, 156, 397
Boruta + JMIM	1, 2, 3, 4, 9, 102
Boruta + IGFS	1, 2, 3, 4, 5, 9, 101, 102, 104
Impurity	101, 102, 103, 106
Boruta + Permutation	101, 103, 105, 106

Best models and the Final Strategy

Selected Feat.	Model	Average Score
101, 103, 104	Naive Bayes	6920
101, 103, 104	Ensemble (NB, QDA)	6940
101, 103, 104	Stacking (NB, QDA + LR)	6950



Conclusions

Conclusions and Findings

- 3 variables suffice as predictors for the classification task
- Naive Bayes high score may signify that the features selected are fairly conditionally independent
- Mutual Information-based methods are slow to compute for too large feature spaces, using Boruta before applying them can drastically improve feature selection analysis time
- Stacked classifier provides stability thanks to utilising two various models

Thank you!