Offer Acceptance Prediction

Project Report

Authors: Izabela Telejko, Bartosz Grabek, Grzegorz Zbrzeżny May 2024

Abstract

This project[1] aims to develop a model to identify customers likely to use the bank's marketing offer based on historical data. We needed to identify one thousand clients out of five thousand test observations who would most benefit from the offer and therefore most likely take it. Due to the high dimensionality of the input training data, a vast range of feature selection techniques was put in place to select a minimal set of predictors yielding the best classification results. The final model evaluation considers both the classification model performance and the number of features used (penalty for each feature is applied). This report describes the most interesting results arising from the most prominent strategies (model and feature selector combinations) that maximize the customized score function. Apart from standalone feature selection methods and classification models, several ensemble methods are considered.

Contents

[1]Source code: https://github.com/Bartosz7/AML-2024L-Offer-Acceptance-Prediction

1 Methodology

To achieve the best results, we conducted numerous experiments, each focused on evaluating a specific strategy, that is a model and feature selector pair. For the experiments, we defined a class called `Experiment` that contained a model and a feature selector along with their configurations. We evaluated each experiment using five-fold cross-validation with `StratifiedKFold` from the `scikit-learn` to ensure similar distribution of the target variable across all folds. For each split we trained our feature selector on the training observations and transformed both the training and validation datasets accordingly. Afterwards, we trained the model and evaluated it on the validation data using the following scoring function adjusted to the cross-validation

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

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.

2 Experiments

In this section, we present the feature selectors and models considered during our experiments. We also highlight the five best-performing combinations and describe which approaches did not yield promising results. Notably, some of the top-performing feature selectors were combinations of two different methods applied sequentially.

2.1 Considered Methods

In our experiments, we used a variety of feature selectors to identify the most relevant features. In some cases, we combined two methods and applied them sequentially, indicated in our results table with a `+` sign between the two method names. We considered the following methods:

- CMIM – Conditional Mutual Information Maximization [1]
- JMIM – Joint Mutual Information Maximization [2]
- IGFS – Information Gain Feature Selection [3]
- Random Forest Feature Importance (Impurity Decrease and Permutation importance)
- Boruta [4]
- Chi-Square Test
- Fisher score

These feature selectors were chosen to provide a diverse set of criteria for feature importance, ranging from statistical tests to model-based importance measures. Moreover, we experimented with a range of models for the binary classification task, including:

- Support Vector Machines (SVMs),

- Logistic Regression,
- Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA),
- Naive Bayes,
- Tree-based methods such as Random Forest, AdaBoost, Gradient Boosting,
- Multi-layer Perceptron,
- Ensembles and stacking of the best models,

These models were selected to cover a variety of algorithmic approaches, from linear models to complex ensemble methods, ensuring a comprehensive evaluation of potential predictive performance.

2.2 Selected Experiment Results

In the Table 1 we present results for the best combinations of feature selector and model with their best configurations. Every experiment configuration can be found by its name in the project repository in `experiment_results`[2] directory. In the Table 2 we presented features commonly chosen by each selection method.

Feat. Sel.	# Feat.	Model	Average Score	Exp. Name
CMIM	4	QDA	5810	qda_cmim_398fe8
	3	SVC	5580	svc_cmim_3f9047
	3	MLP	5490	mlpc_cmim_a19527
Boruta + JMIM	4	QDA	6230	qda_bor_jmim_cebd58
	3	SVC	6040	svc_bor_jmim_dc7f32
	4	MLP	6010	mlpc_bor_jmim_575edc
Boruta + IGFS	3	QDA	6270	qda_bor_igfs_96ad41
	4	SVC	6170	svc_bor_igfs_a208a3
	4	MLP	6130	mlpc_bor_igfs_f18c39
Impurity	4	GaussianNB	6880	gnb_rffis_9ce115
	4	QDA	6850	qda_rffis_838aeb
	4	MLP	6850	mlpc_rffis_905309
Boruta + Permutation	3	GaussianNB	6840	gnb_bor_pi_8902e4
	3	SVC	6800	svc_bor_pi_e42cad
	3	QDA	6830	qda_bor_pi_056032

Table 1: Evaluation of the scores for best considered combinations of feature selectors and models.

[2] https://github.com/Bartosz7/AML-2024L-Offer-Acceptance-Prediction/blob/main/experiment_results

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

Table 2: Features selected throughout the cross-validation splits.

2.3 What did not work?

We attempted to apply various transformations to the initial data and subsequently performed feature selection on the augmented data. However, in most cases, the feature selectors predominantly chose the original features. Even when alternative features were selected, the custom scores did not improve with the transformed features. The transformations we investigated included:

- second-order interactions between the variables,
- mathematical operations such as the square root, square, logarithm, and exponent of the original features.

We applied these transformations to the entire dataset and utilized Random Forest Feature Importance to identify the best features. Alternatively, we applied the transformations to a subset of already selected features and then used the JMIM feature selector to re-select features, but in most cases, this approach resulted in no changes in the selected variables. Moreover, feature selection on the whole dataset using the Chi-Square test or Fisher score did not yield satisfying results.

2.4 Final experiments

After experimenting with various feature selection and binary classification techniques, we achieved satisfying results. However, we wanted to check whether combining features selected in the most promising results might achieve better scores. We tried various combinations of three or four element subsets of features presented in Table 2 and evaluated the results using the best model from previous experiments, namely Naive Bayes. The best score was achieved with features 101, 103, and 104. Next, we explored whether combining a few of the most promising models on the selected features might further increase the score. After conducting experiments with various combinations, we finally created a soft voting ensemble of Naive Bayes and QDA, and a stacking ensemble of Naive Bayes and QDA, with Logistic Regression as the final estimator. The results averaged over cross-validation splits from these final experiments are shown in Table 3. Apart from that in Figure 1 we present boxplots with the results from each of the five cross-validation split.

Selected Feat.	**Model**	**Average Score**	**Exp. Name**
101, 103, 104	Naive Bayes	6920	`gnb_mfs_542793`
101, 103, 104	Ensemble (NB, QDA)	6940	`vc_mfs_c2e45c`
101, 103, 104	Stacking (NB, QDA + LR)	6950	`sc_mfs_7d34e8`

Table 3: Final experiments with best features and models.

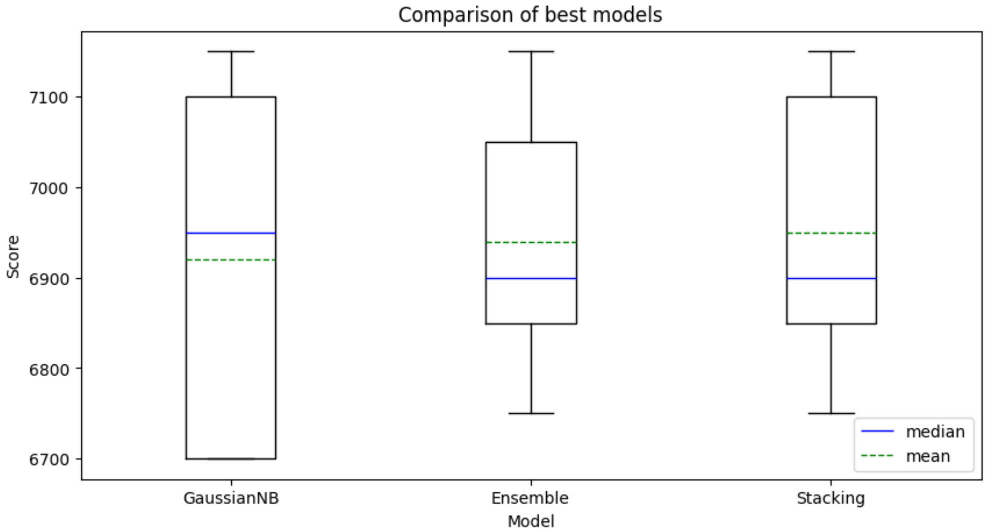

Figure 1: Cross-validation scores of the best models.

3 Conclusions

Throughout the project, our extensive experimentation with various feature selectors and classification models has shown some insightful results regarding the predictive performance of different strategies for the client offer acceptance.

The top performing combinations involved a mix of feature selection algorithms, in particular Boruta algorithm paired with IGFS (or other Information-Theoretic methods). The use of Boruta to decrease the feature space before applying IGSF significantly affected the computation time.

Interestingly, transformations of the features did not help in deriving a better set of features for the classification task. However, feature selectors consistently identified a core subset of features, showing 3 to 4 features as main target predictors. The manually selected subset (features 101, 103, 104) achieved the highest score in our final chosen strategy, which included a stacked classifier utilising Gaussian Naive Bayes and QDA and for the final estimator used Logistic Regression.

The average score of 6950 for the final chosen model is satisfactory, yet there might be still room for improvement. Potential further refinement and deeper hyperparameter optimisation could likely increase the model performance. Also, other more advanced ensemble models could be considered, however, one has to bear in mind that in the real-world scenarios it is often the case that the simpler models are more efficient and easier to maintain. To conclude, the project has been successful as we achieved a satisfactory model at predicting client offer acceptance with just a several features and simple underlying model.

References

[1] F. Fleuret, "Fast binary feature selection with conditional mutual information," *J. Mach. Learn. Res.*, vol. 5, pp. 1531–1555, Dec. 2004, ISSN: 1532-4435.

[2] M. Bennasar, Y. Hicks, and R. Setchi, "Feature selection using joint mutual information maximisation," *Expert Systems with Applications*, vol. 42, no. 22, pp. 8520–8532, 2015, ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2015.07.007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417415004674.

[3] A. El Akadi, A. Ouardighi, and D. Aboutajdine, "A powerful feature selection approach based on mutual information," vol. 8, Jan. 2008.

[4] M. Kursa, A. Jankowski, and W. Rudnicki, "Boruta - a system for feature selection," *Fundam. Inform.*, vol. 101, pp. 271–285, Jan. 2010. DOI: 10.3233/FI-2010-288.