Faculty of Mathematics
and Information Science

WARSAW UNIVERSITY OF TECHNOLOGY

Advanced Machine Learning

Project 2

Jaremek Łukasz, Szymanek Paulina, Wysocka Patrycja

Contents

June 3, 2024

1 Experiments

1.1 Feature selection

The primary objective of this experiments was to identify and select the top most important features from the dataset using chosen feature selection methods. The main requirement is for the model to be parsimonious - to be based on the smallest number of variables, while still achieving the best score.

Considered methods:

- Random Forest feature importance - parameters used were balanced class weight and maximum depth 5,
- Boruta algorithm - algorithm was used combined with Random Forest Classifier with balanced class weight and maximum depth set to 5,
- Least Absolute Shrinkage and Selection Operator (Lasso) regression - training was done on standardized dataset,
- Correlation Feature Selection - the threshold was set to 0.95,
- Genetic algorithm with Gradient Boosting Classifier - GBC with 100 estimators was used. Genetic algorithm was initialized with population of 50 individuals, mutation rate of 10% and 50 generations,
- K-Best method - method was used with the ANOVA F-value as scoring function and selected 10 best features. The data used is the standardized dataset,
- Random search - search based on randomly choosing columns, until scoring function calculating monetary gain specified in the task description achieves value above 7000. It was only used on the smaller column subsets returned by previously mentioned methods.

The methods listed above were used in combinations shown in Table 1 and chosen features shown in the same table. It was done sequentially and based on that the best columns were chosen. The last two mentioned methods were only used in the experimental stage.

Features were also analyzed based on correlation between them. Correlation plot for features, where any of the features has correlation above 0.4 threshold is shown in Figure 1. The most correlated features are features number 0 to 9 with correlation higher that 0.7. The same features are slightly correlated with features 199 to 109, with correlation around 0.25.

Method	Selected Features
Boruta + RFE + Correlation	8, 100, 101, 102, 105
RFE + Boruta + Correlation	100, 101, 102, 103, 105
RFE + Correlation	100, 101, 102, 103, 105
Boruta + Correlation	0, 1, 2, 3, 4, 5, 7, 8, 9, 100, 101, 102, 103, 104, 105
Boruta + Lasso	105
Boruta + RFE + Lasso	105

Table 1: Features selected by different feature selection methods

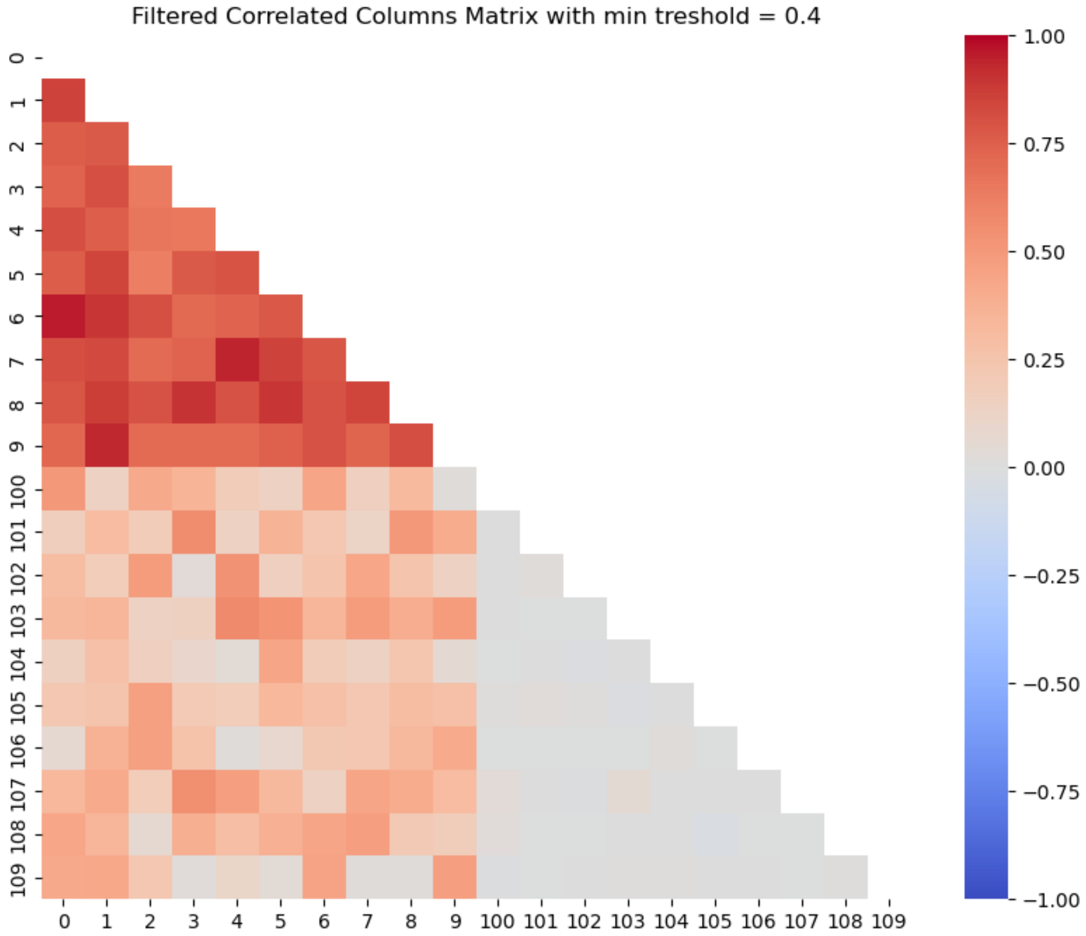

Figure 1: Correlation plot of features

1.2 Models

This analysis involves evaluating several machine learning models for binary classification tasks. The data used includes a feature set selected with feature selection methods described in 1.1.

Considered models:

- Gaussian Naive Bayes (GaussianNB) - model was used with no parameters set,
- Random Forest Classifier - model was used with number of estimators set to to 59, maximum, depth of 30, maximum features 0.999, minimum samples leaf 1, minimum samples split of 10 and number of estimators of 1000,
- XGBoost with Random Forest (XGBRFClassifier) - maximum depth was set to 17, learning rate set to 0.9, number of estimators set to 80, subsample of 0.5 and evaluation metric set to logarithmic loss,
- XGBoost - maximum depth was set to 50, learning rate set to 0.007, number of estimators set to 100, subsample of 0.1 and evaluation metric set to logarithmic loss,
- Logistic Regression - model was used with l2 penalty norm, inverse of regularization strength equal to 1.0 and lbfgs solver,
- Multilayer perceptron (MLPClassifier) - activation function was set to tanh, alpha to 0.007, hidden layer sizes to 98, learning rate to 0.02 and solver to sgd.

Best parameters for the last five models were chosen using Bayesian optimization over hyper parameters. It was done with k-fold cross validation with 5 splits and recall as the scorer. Model training was done using k-fold cross validation with 10 splits and model evaluation was done using cross validation with recall as the scorer.

The best models were chosen by selecting columns with numbers 8 and from 100 to 105 and then running models on different combinations of those columns. Each experiment was done on multiple train-test splits. Final random search was done on selected columns with monetary score as the metric. Monetary score can be defined as:

$$monetary\_score = num\_correct * 10 - num\_variables * 200, \tag{1}$$

where *num_correct* is the number of correctly predicted observations of target variable and *num_variables* is the number of variables used in model training.

2 Results and conclusions

Table 2 shows the best achieved performance of the models and the columns that were used. The final model chosen for the prediction was Gaussian Naive Bayes used on columns with numbers 101, 102, 103 and 105, as it achieved the best score overall. The model was run 1000 times, its results averaged and it always achieved score above 7000.

Model	**Columns**	**Score**	**Recall**
Gaussian Naive Bayes	101, 102, 103, 105	7075	0.58
Random Forest Classifier	102, 103, 105	6600	0.56
XGBoost with Random Forest	101, 102, 103	6640	0.55
XGBoost	102, 103, 105	6730	0.6
Multilayer perceptron	101, 102, 103, 105	6882	0.65

Table 2: Performance of different models

Based on multiple runs on different combination of columns, best results in case of monetary score were achieved with Gaussian Naive Bayes model, while still maintaining better or comparable recall than methods such as Random Forest Classifier. Multilayer perceptron was able to achieve the highest recall on various combinations of columns and achieved better monetary score than all of the other models except for Naive Bayes. Better results may have been obtained by using different hyperparameters, hyperparameter search method, additional classification models or using ensemble of chosen models.

The number of initial variables in the training data made the feature selection task difficult. The small amount of observations added to the challenge and complicated model validation process.