GitHub: https://github.com/mateschieszler/DA3-A3

Assignment III

Data Analysis 3

Máté Schieszler, Péter Szilvási

2024.03.03

Introduction

The goal of the project is to build the best possible model to predict company defaults in the 'Manufacture of computer, electronic and optical products' industry in the year 2015.

The data

The dataset contains detailed company data from a middle-sized country in the European Union. The database was constructed for from multiple publicly available sources by Bisnode, a business data and analytics company since then acquired by Dun & Bradstreet. The dataset contains 287,829 observations on 46,412 firms from 2005 to 2016 (the data from 2016 only contains third as many companies as in previous years). The id variables are company identifier (comp_id) and year (year).

The race

Half of the maximum 30 points for the assignment are allocated according to model performance. The goal is to achieve the lowest possible loss value on the hold-out sample. The hold-out is 1037 SMEs in 2014 in the chosen industry ('ind2 = 26'), out of which 56 defaulted and 981 stayed alive. Further info on the firms: average sales is 0.4902 million EUR, with the minimum of 0.00107 million EUR and the maximum 9.57648 million EUR. Using this sample in any way to train the model should be avoided and is penalized by a 10-point reduction.

Data preparation

The first and most important step of the data preparation is determining which companies defaulted, in other words, the creation of the 'default' variable. We define this as companies that had positive sales in a given year and had no sales or are missing from the dataset in the next year. The EDA process is detailed in the technical report. During further preparation we keep the holdout set and its defined description metrics unchanged. The variables with high ratio of missing values are dropped. We create new numerical and categorical variables describing firm characteristics. To keep the holdout set the same, as in the assignment description, we impute important but missing variables, where it only concerns a small share of records. Flags are used for signalling imputation, and possible problems with

engineered variables (too low, too high, below zero, where it shouldn't). Certain columns, such as the growth measured in difference in sales ('d1_sales_mil_log') are winsorized at the 90th percentile.

Modelling

We create 4 models to predict defaulting companies and choose the best performing for the final prediction. We compare them based on cross validated average loss on the training data.

Training data

We concluded that filtering for the chosen industry ('ind2 == 26') creates the best training sample, compared to the total panel data, or looking at the cross section for the chosen year ('year == 2014'). This confirms that there is a low external validity between industries in terms of defaults. Filtering this way the training data has 10,055 records on 1604 firms.

OLS

We build five increasingly complex OLS models including more and more variables from X1 to X4 and also interactions for X5. Out of the five X4 performs the best so we choose this for representing the OLS models. The best performing OLS model has an 0.244 average RMSE and an average loss of 0.833 on the training set.

LASSO

We also run a LASSO algorithm on the most complex OLS model (X5), greatly reduces the number of variables from 258 to 188. The LASSO model has a cross-validated average RMSE of 0.246 and loss of 0.848 on the training set.

Random Forest

We define the variables for random forest separately as not to include interactions and variables which are derived from multiplication or quadratic functions from each other. The optimal random forest has a maximum of 6 features in each split and a minimum of 9 elements in each split, when examined by the Gini index criterion. The Random Forest model has a cv average RMSE of 0.242 and a loss of 0.822.

GBM

To improve the Random Forest model with use Gradient Method Boosting. Using grid search we determine that the best performing model has a maximum depth of 13, maximum number of features 6, and 300 estimators. While the GBM creates a better cv average loss on the train set as the random forest, both the cv RMSE and the cv AUC lags behind.

GitHub: https://github.com/mateschieszler/DA3-A3

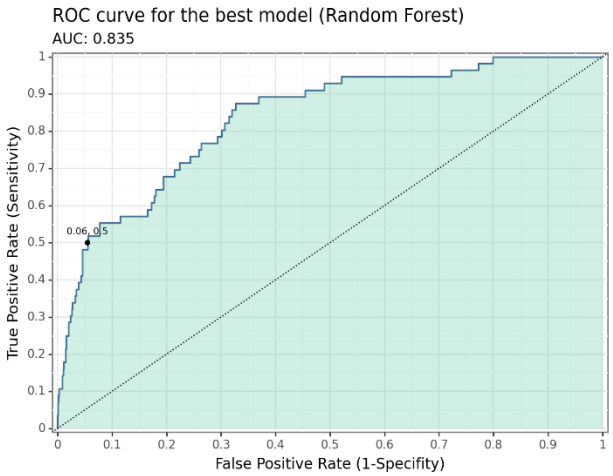
Final Prediction

The performance of the models on the training set can be summarized in the following table:

MODEL	PREDICTORS	CV RMSE	CV AUC	CV THRESHOLD	CV EXPECTED LOSS
X4	167	0.2444	0.7867	0.1575	0.8327
LASSO	188	0.2461	0.7754	0.1642	0.8476
RF_P	65	0.2418	0.8132	0.1826	0.8219
GBM_P	65	0.2605	0.7922	0.0007	0.8172

For the final prediction of defaults we use the random forest model, as the model with the best overall measures. When used on the hold-out set the model performs with the following measure and produces the following confusion matrix.

CONF. MATRIX	ACTUAL NO DEFAULT	ACTUAL DEFAULT
PREDICTED NO DEFAULT	927	28
Predicted default	54	28



Measure	VALUE
BRIER-SCORE / RMSE	0.2093
AUC	0.8353
Best Threshold	0.1825
Accuracy	92.093%
Sensitivity	50%
Specificity	94.495%
Loss	0.561

The final loss value on the hold-out is **0.561,** while the hold-out set remains unchanged in terms of the descriptive statistics in the assignment description: The hold-out is 1037 SMEs in 2014 in the chosen industry ('ind2 = 26'), out of which 56 defaulted and 981 stayed alive, with average sales of 0.4902 million EUR, with the minimum of 0.00107 million EUR and the maximum 9.57648 million EUR.