Prediction ML 2023

Assignment 3

Emil Savalanlı

Motivation

This task involves solving the business problem of identifying the best firm (i.e, fast-growing) to invest in for two-year period using available data from the Bisnode database. I analyze this using firm data from 2010 to 2012 to predict their growth in 2013 and 2014. I built several models.

Sample

Overall there are more than 300k observations in the given panel data that covers 2010-2015 years. I dropped firms with negative sales from 2010-2014, because those years are ones for the analysis and by doing this to some extent I am cutting out fat tails that could spoil model predictions. After some cleaning I ended up with around 19k observations in 2010-2012 sample, from this 15k is being used for training of the models.

Target variable

A choice of the target variable is an important issue. But to somehow represent "investing" environment or picture I am using ROE to construct target variable. However, it is not an exact ROE. To represent two-year investing picture, I constructed pseudo-return variable from ROE, it is basically compound return for the two years calculated using ROE for 2013 and 2014. Thus, first, for example, using sales growth as a target variable that takes a difference between two years, I think, causes information lost. But with my proposed target variable I am able to use information of 2013 and 2014, second, as already speculated, ROE is closer to cases in developed markets where investors do care about equity, shares, stock price etc. more than about just total assets. Additionally, in developed markets higher than 15% considered as a one of fast-growing firm, but to be robust, I take 20%, thus two-year compound ROE for fast-growing firms is around 50%. Hence, the final target variable is equal to 1 if compound ROE is higher than 50%, and 0 otherwise.

Feature engineering

Firstly, I dropped the features(columns) that have missing values more than 30% of whole data set observations. Because almost one-third is missing, I think, those features is going to be unable to generate any variations in the models. I created additional features: they are financial ratios, namely, liquidity ratio, turn over ratio, ROA, ROE.

For NAs (missing values) I mainly used a median of the respective columns to fill in. I flagged imputed values. In initial analysis I dropped some features after exploring a heat map of the cross-correlation matrix. Note that the matrices are separately explored for analysis years (2010, 2011, 2012) to avoid serial(auto)-correlation. Hence, I dropped variables as tangible assets, subscribed cap, material expenditure, extra income, inventories, some industry codes features and fixed assets. I created dummy variables by myself from nace_main feature by extracting main categories of the industry areas. Missing values for main financial variables and some firm values are filled in with 0 and flagged.

I take square of the values of numerical columns to account for some non-itineraries. I take the log of sales, and further I log-ed all numerical columns that contains values larger than 10 and less than -10. By doing that I standardizing or normalizing the data set, logs are helpful with outliers, additionally it is to prevent blowing off of the models, because the numbers are larger, not to mention squared values of them. Positive and negative infinity values are converted to the minimum of maximum of the respective column's values.

Models

Train and test set are split as 0.8/0.2. Because of CV usage test set there is as hold-out set. The models cross validated with 5 folds on the train set. To check models external validity in terms of live data one could predict growth of firms in 2015 by using data from 2014, 2013 and 2012. But below I show performance of models on test(hold-out) set

Loss function

Actually I am using profit function as a loss function. Anyway, it could be converted to minimizing problem by multiplying by -1. But I go with profit function and maximizing it over thresholds. Basically, it is

$$1.5 \cdot TP - 1.5 \cdot TN - 3 \cdot FP$$

The intuition is that the fast-growing firm returns +50%, so that is TP, but one couldn't invest in TN firms, that's actually fast-growing but predicted as not, so lost on it is -50% of TN firms. And I harshly penalize as double of average return from fast-growing firm, 3 times FP. It could be explained as a leverage,

so, arbitrary investor loses more on bad choices than gaining on good ones, as in usual markets.

Logistic Regression

Lasso and Ridge regularization give the same performance, so I just take Ridge one. F1-score is 76% and AUC is 82.6%. Optimizing by F1-score gives optimal threshold of .36. But with my defined loss function it is around 0.5

Random Forest

This also gives F1-score around 76% and AUC of 82.6%.

XGBoost

This model performed with F1-score of 82.7% and AUC of 83.2%. Optimal threshold is way higher than in case of logistic regression. Perhaps it is problematic

Catboost

This model performed with F1-score of 73.5% and AUC of 83.3%.

Overall all forest-boosting models above showed optimal classification threshold of 0.89. That's logical. Because one wants to predict as lower as possible number of false positive, so it covers right bottom part of ROC-AUC curve

KNN

This model performed with F1-score of 73.5% and AUC of 74%

Technical report

Can be found here