Macroeconomic forecasting: Can machine learning methods outperform traditional approaches?

CFDS06 Project

Felix Jobson

29.05.2021

Table of contents

Problem Description

Problem Description

Financial Data Science

- The research question of the project is the capability of machine learning models to predict the growth of an economy and compare the result with traditional methods of forecasting.
- The dependent variable is the growth rate of the gross domestic product (GDP). This is the objective of the learning and prediction task. The independent variables are several macroeconomic factors.
- The baseline models are classical econometric methods and the World Economic Outlook of the International Monetary Fund.

Data

Data sources

Financial Data Science

- Sources
  - International Monetary Fund (IMF)
  - Organisation for Economic Co-operation and Development (OECD)
- Time Period: 1980 - 2017
  - Training Set: 1980 - 2004
  - Validation Set: 2005 - 2010
  - Test Set: 2011 - 2017
- Countries:
  - Initially 189
  - After clearning 46

Variables

Financial Data Science

- Number of macroeconomic factors used:
  - Initial: 41
  - After cleaning 15
- Examples of used varibles
  - Inflation
  - Unemployment rate
  - Material consumption
  - Working age population
  - Fertility rates

Split of the dataset

Financial Data Science

- Two different purposes:
- Model selection and model assessment.
- The validation set is used to estimate the prediction error for model selection.
- The test set should be kept in a "vault" and is used to estimte the test error at the end of the analysis.

Transforming to Growth Rates

Financial Data Science

- Because the variable have different absolute values, growth rates are used.
- To receive the same magnitude for an increase as well as a decrease a logarithmic transformation is used:
- $$\hat{x}_i = \ln(\frac{x_i}{x_{i-1}} + |\min_j(x_j)| + 0.001)$$

Preparation of the Data

- Using the framework of supervised learning to work with time series.
- The orignal data is given in the form (x_t, y_t), $t = 1...N$
- For every time step the outcome y is mapped to predictor variables x that are preceeding:

$$(x_{t-1}, y_t), \ t = 2...N$$

- Hence a model for supervised learning can be trained and used for prediction.

Impute missing values

Financial Data Science

- Only countries with less than 50 % missing values are used. Then the top 15 filled variables are selected.
- To use time series with missing data at all, an imputing strategy is used: *k-nearest neighbors*
- Each sample's missing values are imputed using the mean value from n nearest neighbors found in the training set.
- Important: Fit on the training set and then apply imputation on the validation and test set.

Approach

World Economic Outlook

Financial Data Science

- The International Monetary Fund publishes predictions of the GDP growth in its World Economic Outlook (WEO)
- The IMF publishes the WEO twice a year in spring and fall.
- The prediction from the fall is used, as this is closer to the next year and therefore the prediction is more precise.

Classic Models

- Ordinary Least Squares
  - The OLS regression is the most famous and basic model in econometrics. It has the following form:

$$y = x_1\beta_1 + x_2\beta_2 + ...x_N\beta_N + \beta_{N+1}$$

- Autoregressive Integrated Moving Average
  - The autoregressive integrated moving average ARIMA(p, d, q) model is used in time series analysis.
  - $X_t - \alpha_1 X_{t-1} - ... - \alpha_p X_{t-p} = \epsilon_t + \theta_1 \epsilon_{t-1} + ... + \theta_q \epsilon_{t-q}$
  - Here α_i are the parameters of the autoregressive part of the model, θ_i are the parameters of the moving average part, d is the degree of differencing and ϵ_t are error terms.

- Least Absolute Shrinkage and Selection Operator
  - The LASSO is a penalized version of the OLS:

$$\min_{\beta} ||X\beta - y||_2^2 + \alpha||\beta||_1$$

- Support Vector Regression
  - The SVR is an adapted version of a SVM for regression problems and tries to solve the optimization problem:

$$\min_{\beta} \frac{1}{2}||\beta||_2^2, \text{ subject to } ||X\beta - y|| < \varepsilon$$

- Regression Tree
  - Binary tree that groups data with similar vaules into the same leaf. The response in each leaf $L_1, L_2, ..., L_M$ is modeled as constant, so the tree can be expressed as a function:

$$f(x) = \sum_{i=1}^{M} c_m I(x \in L_m)$$

- Gradient Booster
  - Ensemble of the from

$$f(x) = \sum_{i=1}^{N} f_i(x)$$

    where f_i are weak learners, most of the time tree based models.
  - Are called gradient booster because of the way the model is trained.

- Recurrent Neural Network
  - A RNN is a deep neural network that is designed to handle sequential data.
  - A RNN cell is defined as:

$$h_t = \sigma(W_{ih}x_t + b_{ih} + W_{hh}h_{(t-1)} + b_{hh})$$

  - There are also more sophisticated approches like the LSTM (Long short-term memory).

Evaluation

- The performance of the models is measured by the MSE of the test set:

$$MSE = \frac{1}{|T|} \sum_{t \in T} (y_t - \hat{y}_t)^2$$

- The cartesian product Ω of the set of all classical models and all machine learning models is formed. The MSE of both is compared:

$$X(\omega) := \begin{cases} 1 & \text{if } MSE_{ML_\omega} > MSE_{classic_\omega} \\ 0 & \text{else} \end{cases} \qquad \omega \in \Omega$$

Evaluation

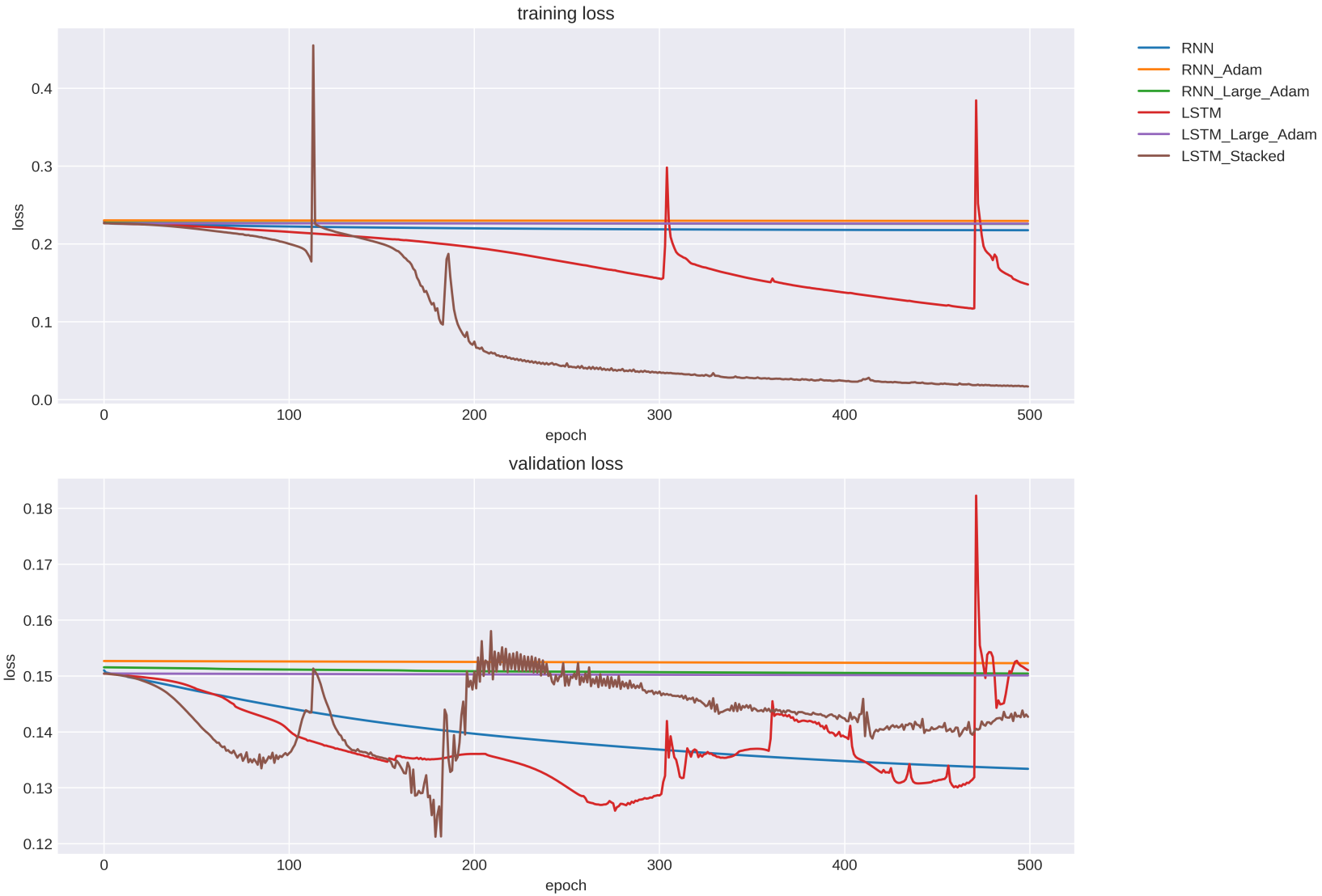
Financial Data Science

- Confidence intervals of X are approximated by bootstrapping.
- If the lower bound of the confidence interval is greater than 0.5, the machine learning methods have statistically significant better performance than the traditional approaches.
- Evaluating this approach with two different settings:
- Training each model with the data of a single country.
- Training each model with the whole data combined.

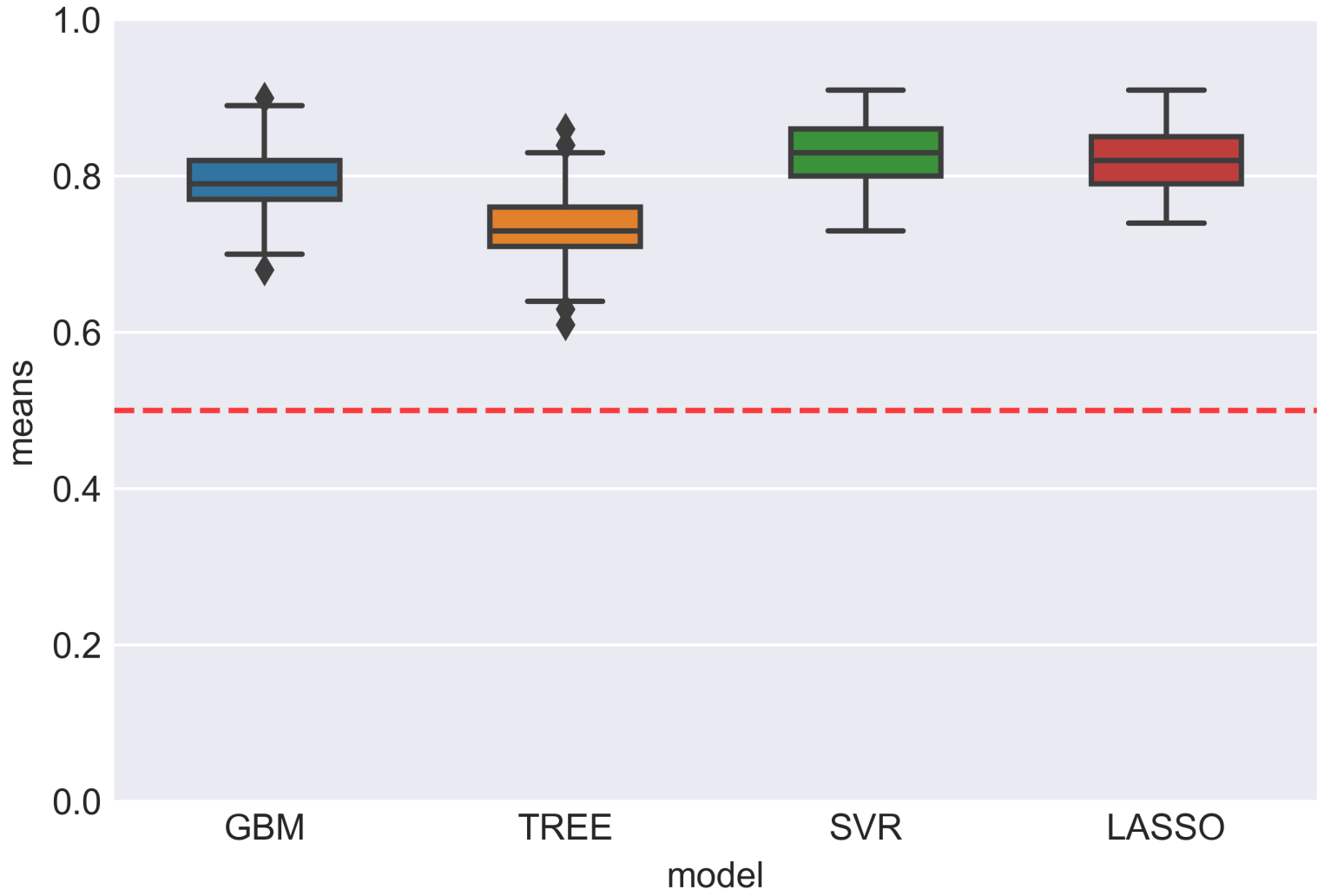
Results

Deep Learning went wrong

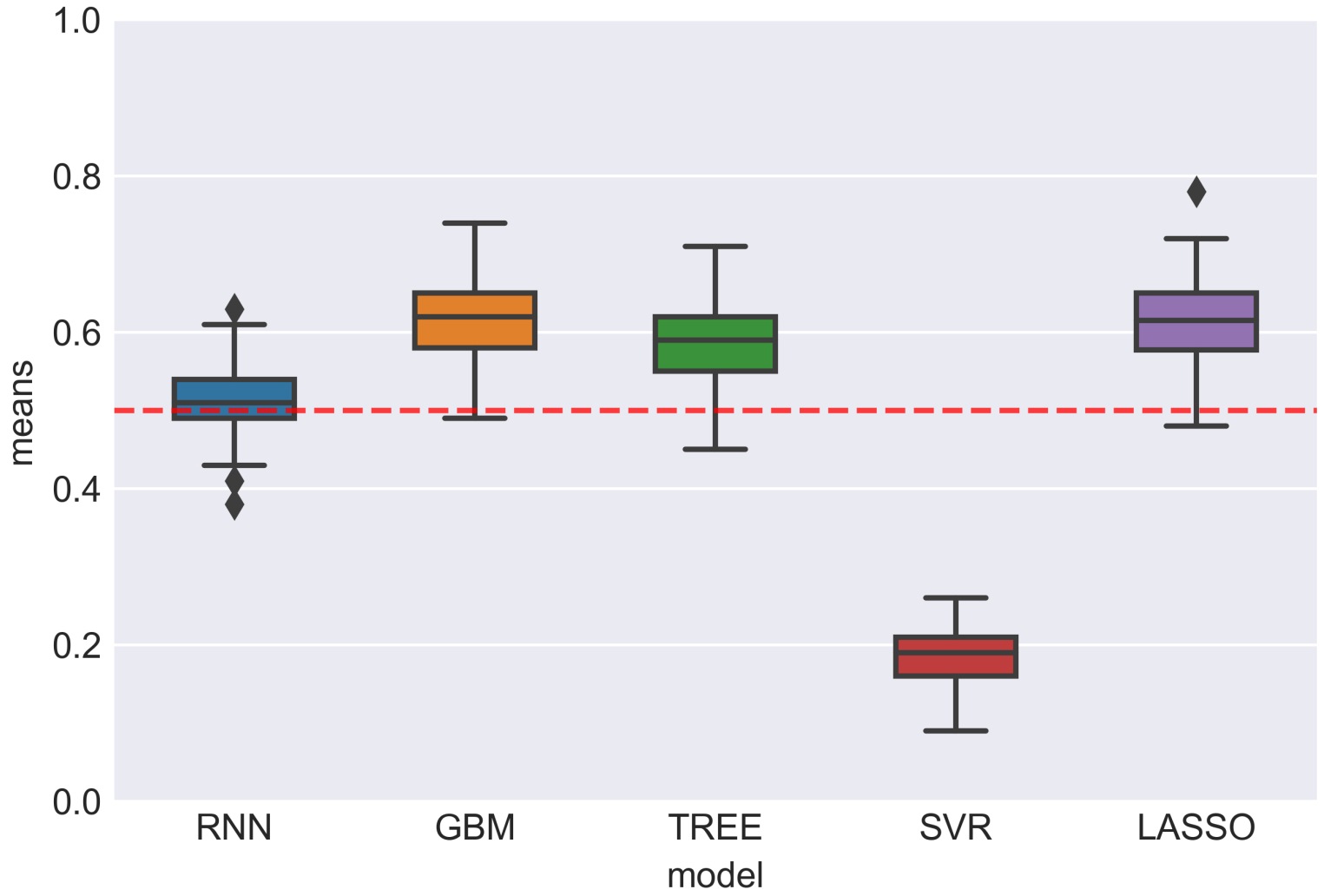
Financial Data Science



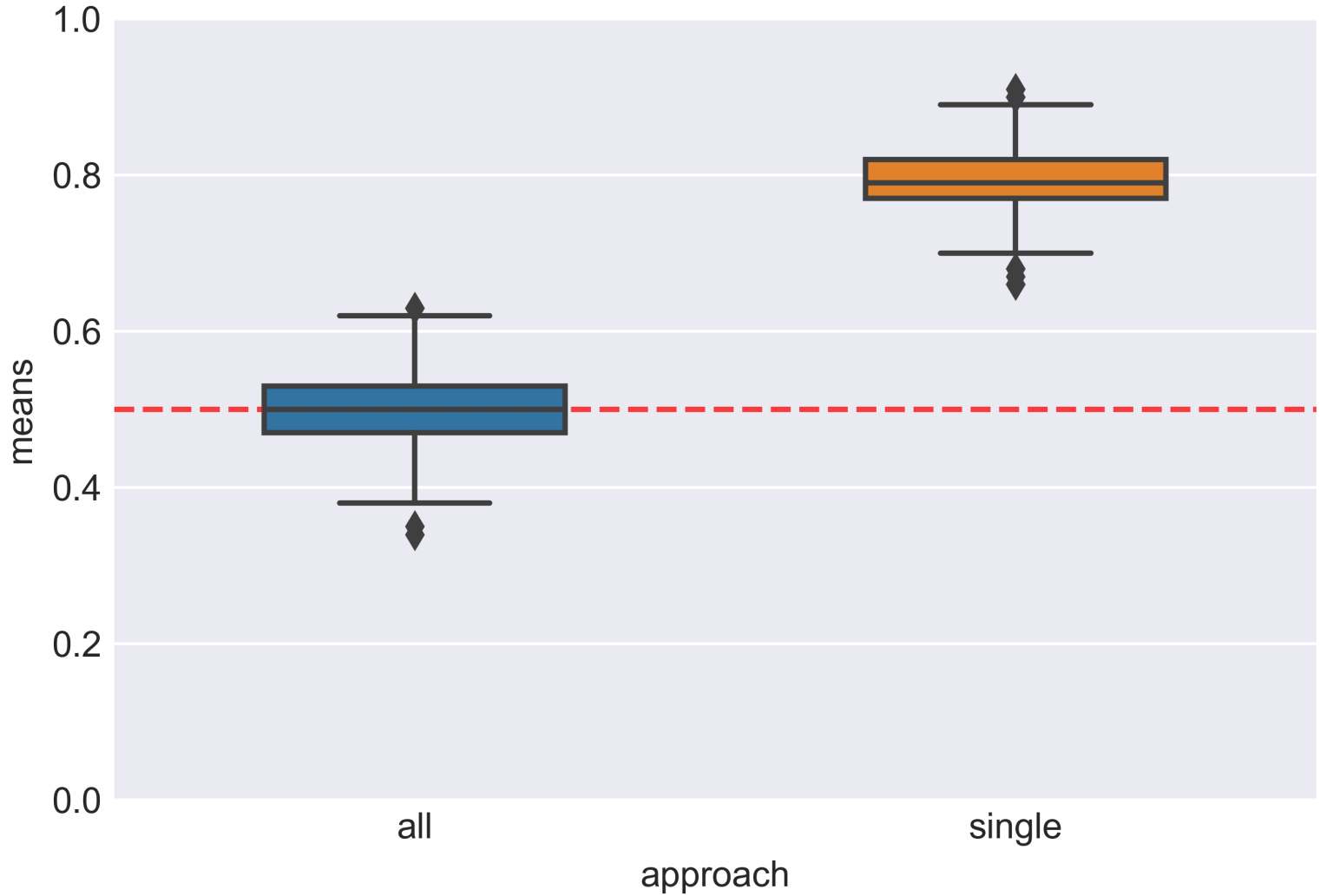
Result Training Single Country



Result Training All Countries



Test for statistical significance



Conclusion

Conclusion I

Financial Data Science

- Machine learning models can outperform traditional approaches!
- At least in the given evaluation framework presented.
- Data collection and handling take the most time from the project budget, modelling takes only a fraction.
- Deep learning relies heavily on the amount of data and fails if there is not enough available.

Conclusion II

Financial Data Science

- Even simple machine learning models have a decent performance.
- SVR failed on training with all countries. A profound understanding of the model is important to understand problems.
- The proposed deep reinforcement learning approach was not successful.

Possible Next Steps

Financial Data Science

- Collect better Data in terms of quality and quantity.
- The "expert-based" decision should be derived based on data.
- Analyse feature importance and automate feature selection.
- Analyse the transformation of the data and use a more sophisticated approach.

"All models are wrong, but some are useful"