Explaining Multivariate Time Series Forecasts: an Application to Predicting the Swedish GDP*

Henrik Boström[1], Peter Höglund[2], Sven-Olof Junker[2], Ann-Sofie Öberg[2], and Martin Sparr[2]

[1] KTH Royal Institue of Technology
bostromh@kth.se
[2] The Swedish National Financial Management Authority
{peter.hoglund,svenne.junker,ann-sofie.oberg,martin.sparr}@esv.se

Abstract. Various approaches to explaining predictions of black box models have been proposed, including model-agnostic techniques that measure feature importance (or effect) by presenting modified test instances to the underlying black-box model. These modifications rely on choosing feature values from the complete range of observed values. However, when applying machine learning algorithms to the task of forecasting from multivariate time-series, it is suggested that the temporal aspect should be taken into account when analyzing the feature effect. A modification of individual conditional expectation (ICE) plots is proposed, called ICE-T plots, which displays the prediction change for temporally ordered feature values. Results are presented from a case study on predicting the Swedish gross domestic product (GDP) based on a comprehensive set of indicator and prognostic variables. The effect of calculating feature effect with and without temporal constraints is demonstrated, as well as the impact of transformations and forecast horizons on what features are found to have a large effect, and the use of ICE-T plots as a complement to ICE plots.

Keywords: Explainability · Forecasting · Multivariate time series · GDP

1 Introduction

Machine learning for time series analysis has received significant attention over the years, in particular classification of (univariate or multivariate) series, see e.g., [8]. The task of forecasting, i.e., predicting how the time series will extend beyond the latest observed time point, rather than labeling the time series, has also received some attention within the machine learning community, with work in the area dating back several decades, see e.g., [4].

In a recent study [7], researchers at the International Monetary Fund (IMF) investigated the use of machine learning for multivariate time series forecasting of the gross domestic product (GDP) one quarter and four quarters of a year

ahead, respectively, for a number of countries, where macroeconomical variables and previous outcome (GDP) observed over a few decades were used to predict the future GDP change. It was found that the machine learning models not only outperformed traditional statistical techniques, but also IMF's own World Economic Outlook (WEO) forecasts. However, the model with the strongest predictive performance consisted of an ensemble produced by the Super Learner [9], i.e., effectively a black-box model, and in the discussion of future work, the researchers consequently pointed out the need for methods that are able "to unbox and interpret machine learning models to provide explanation for their outputs, and help understand the differences in forecast performance across a wide-range of model and expert-based forecasts".

In this study, we will consider the task of forecasting the GDP of Sweden, a country not included in IMF's original study, and investigate the application of techniques for explaining predictions, by analyzing the impact each feature has on predictions of the underlying (black-box) model. In the standard machine learning setting, where a model is trained from a set of examples that is assumed to be independently sampled according to some fixed but unknown distribution, the effect (or importance) of a feature with respect to the model, may be estimated by measuring how the predictive performance (or the predictions) change when modifying the values of the features for a set of examples. Such a procedure was proposed in [3], which aimed for explaining random forests by providing estimates of the *variable importance*, measured by the performance degradation when randomly permuting the values for each feature in turn. The procedure exploited the fact that the performance of random forests (like any model generated by bagging) can be estimated using out-of-bag predictions, but the procedure may be straightforwardly applied also to other models, if a separate dataset is used to measure performance degradation. Another procedure for investigating the impact of features on the model is the *partial dependency plot* (PDP), which was proposed in [5]. Rather than measuring the effect on predictive performance, such a plot shows the output of the (black-box) model for possible values of a selected feature (or subset of features), averaged over a sample of examples for which values for the selected feature are varied, while keeping the values for the non-selected features unchanged. It should be noted that both variable importance and PDPs, as calculated by the above procedures, estimate the impact on a set of examples, rather than estimating the effect for a specific prediction. However, in a typical forecasting scenario, we will update or generate a new model for each new observation in the series, and there will hence not be a single (black-box) model used to make the predictions, but a sequence of models. Moreover, we are not mainly interested in the general properties of these models, but rather in features affecting the specific prediction for which each model is used.[3] Hence, rather than trying to characterize global properties

[3] Each model in the sequence is assumed to be used for making only one prediction, relative to some specified time-point, i.e., the forecast horizon. In case predictions for multiple future time-points are needed, a separate model is assumed for each horizon.

of each model, we are interested in properties relating to the specific prediction, for which the model is used. The *individual conditional expectation* (ICE) plot [6], is an adaptation of PDP for individual examples, which instead of averaging over a set of examples, calculates and displays the effect of individual feature values on the resulting prediction for a specific instance. However, similar to PDPs, ICE plots display the predictions as a graph over all possible feature values for the selected feature(s) vs. the resulting prediction, and applying this directly to forecasting models trained from time-series means that the temporal dimension is lost. In this work, we propose a complementary visualization, called the ICE Temporal (ICE-T) plot, which displays the prediction changes for temporally ordered feature values.

In the next section, we describe the various approaches to explaining predictions by feature effects in more detail, including the novel ICE-T plot. In Section 3, we describe the considered prediction task and dataset together with the experimental setup and present findings from the empirical investigation. Finally, in Section 5, we summarize the main conclusions and outline directions for future research.

2 Calculating Feature effect

In the following, we will assume that we have an ordered set $X \in \mathbb{R}^{N \times D}$ of N objects with D features, and an ordered set of labels $Y \in \mathbb{R}^N$. Let $X_{i,j}$ denote the element at row i and column j in the matrix X, and Y_i denote the ith element of the vector Y. Let $X_{i,:} = (X_{i,1}, \ldots, X_{i,D})$ denote the ith row (object) of X, $X_{a:b}$ denote the sequence of rows $(X_{a,:}, \ldots, X_{b,:})$, $X_{:,i} = (X_{1,i}, \ldots, X_{N,i})$ denote the ith column of X, and $Y_{a:b}$ denote the sequence of elements $(Y_a, \ldots, Y_b)$. Let $u(x = (x_1, \ldots, x_D), i, v) = x' = (x_1, \ldots, x_{i-1}, v, x_{i+1}, \ldots, x_D)$, i.e., given an object $x \in \mathbb{R}^D$, the function returns an updated object $x' \in \mathbb{R}^D$ where the ith feature value x_i has been replaced by the value v.

We will moreover assume that we have an underlying (black-box) model M, such that given an object $x \in \mathbb{R}^D$, it returns a (predicted) label $\hat{y} = M(x) \in \mathbb{R}$.

Given an object $x \in \mathbb{R}^D$, a set of values V and a model M, the *feature effect* FE on feature i, is defined as:

$$FE(x, M, i, V) = M(x) - \frac{\sum_{v \in V} M(u(x, i, v))}{|V|} \tag{1}$$

The above function hence calculates the difference between the original prediction for the object x and the average prediction from updating the object on feature i with values from V.

In case $V = X_{:,i}$ for some random sample X drawn independently from the same distribution as the training set that was used to construct M, then FE provides an estimate of the expected prediction change relative to this distribution (and feature). However, as we are here primarily interested in understanding what effect different features have on a specific prediction, rather than providing an unbiased estimate of the expected change for unseen examples, we will,

as commonly done, allow the feature effect to be estimated with respect to the training (or any other) set. It should be noted that the feature effect is in contrast to variable (or feature) importance, not defined in relation to the labels (Y), but only considers the average change of the prediction.

When producing PDP [5] plots, the predictions are averaged over multiple objects, which in turn are updated with respect the full range of possible values for the feature (as observed in the training set). In contrast, an ICE plot [6] is calculated with respect to a single object, again using the full range of observed values. Given an object $x \in \mathbb{R}^D$, a set of objects $X \in \mathbb{R}^{N \times D}$, an underlying model M and a feature j, as defined above, an ICE plot can be defined by the following set of points:

$$ICE(x, M, i, X) = \{(v, FE(x, M, i, \{v\})) : v \in X_{:,i}\} \tag{2}$$

When plotted, with the feature values on the x-axis and the feature effect on the y-axis, the points are normally connected by lines, effectively providing an interpolation of the predictions of the underlying model between each pair of consecutive feature values.

For multivariate time series data, we typically have timestamps for the objects in X, i.e., $T = (t_1, \ldots, t_N)$. Assuming these time points to be unique, we define an ICE-T plot by the following set of points:

$$ICE\text{-}T(x, M, i, X, T) = \{(t_h, FE(x, M, i, \{X_{h,i}\})) : h = 1, \ldots, N\} \tag{3}$$

In contrast to the ICE plot, the ICE-T plot hence allows for visualizing the feature effect over time, where the x-axis represents the time at which feature values have been observed, rather than specific feature values. Since the actual feature values are not included in an ICE-T plot, it should be considered to give a complementary view to the ICE plot.

3 Empirical Investigation

In this section, we first describe the prediction task that is considered in this study. We then describe how the empirical investigation has been designed, including the data preparation, before presenting the findings from the empirical investigation.

3.1 Swedish GDP data

GDP, or gross domestic product, is a measure of the total economic activity taking place on an economic territory which leads to output meeting the final demands of the economy. GDP is an aggregate in the national accounts, an accounting system meant to summarise and describe the country's economic activities and development. There are in principle three ways to compute GDP:
1) *the production approach*, which is the sum of all value added from produced

goods and services, 2) *the expenditure approach*, which is the sum of all expenditures made for consuming the output of the economy or adding to wealth, or 3) *the income approach*, which is the sum of all incomes earned by producing goods and services [2]. There is therefore a number of components that together form GDP or other aggregates of economic activity. The data from the National accounts in this model primarily concern the second approach to GDP, i.e., the expenditure approach. Thus, the model is dependent on levels of household and government consumption, investments and exports and imports.

Swedish GDP is compiled and published by Statistics Sweden (SCB). GDP and other national accounts are made in accordance with a common European standard [2]. The GDP for previous periods are constantly revised. At each quarterly publication there are usually revisions for the latest previously published quarterly numbers. There is also a general revision each year, which could lead to revisions up to 25 years back in time.

In addition to parts and aggregates of the national accounts, there are also other economic features that should have predictive power for estimates of GDP. For this study, we have chosen the most prominent features in the economic outlook in forecasts from the Swedish National Financial Management Authority. Examples of such features are unemployment figures, inflation and interest rates. Apart from Swedish domestic economic features, there are also exchange rates and foreign interest rates as those are relevant for the export-oriented economy of Sweden.

There are also a small number of economic indicators used as features. These can be backward- or forward-looking (GDP in itself is often used as a backward-looking indicator of the general health of the economy). Backward-looking indicators that are used in this model are for example reported vacancies from employers, redundancy notices and number of newly purchased cars. Indicators like these are often used as they convey trends in actual, not calculated, economic activity. They are used both in forecasting and in business reporting in the general news.

Forward-looking features are in general used to capture intent and predictions for the future through surveys. Answers from such surveys are then summarised to an index which can be tracked over time. An example of such a study incorporated in this model is the Swedish Economic Tendency Survey conducted by the Swedish National Institute for Economic Research. The results from the survey is used to construct an indicator. The Economic Tendency Indicator is based on monthly surveys of households and firms and consequently captures the sentiment among these players in the Swedish economy. The indicator is based on the information contained in the confidence indicators for manufacturing, services, construction, retail and consumers [1].

All features used in the model have values on a quarterly basis. The values for some of the features are available on a monthly or even daily basis, but they are either summed or averaged to a quarterly value in the model. Data is available for most features from 1993 up to the last quarter of 2018, and features with missing values in this time span have been excluded. The complete multivariate

time-series considered in this study contains 104 objects with 68 features, in addition to time points and outcome (GDP).

3.2 Experimental setup

The experiment has been designed to emulate a realistic scenario, where only data available up to the point in time in which a prediction is made, may be used for generating (and explaining predictions of) a model to predict the outcome (GDP) at a specified later time point. We here consider two time frames; predicting the GDP one quarter of a year ahead and four quarters of a year ahead, respectively. Moreover, in addition to considering the actual GDP as the target variable, we will consider predicting the percentage change relative to the time point at which the prediction is made, and investigate how this transformation affects both predictive performance and feature effects. In addition to using the features described in the previous section, we will also, as is common in time-series forecasting, use the (recent) previous outcomes as features; in this study we will use five features to represent the outcome of the current (the time point at which the prediction is made) and four preceding quarters, which we refer to as *lagged values*. It should be noted that this means that some of the lagged values will be missing for the first four objects in the time series (as we do not know what the earlier outcome was), and rather than handling missing values, e.g., by imputation, we simply exclude these objects from the training set. Furthermore, if we at the current time point t_h want to make a prediction for time point t_{h+a}, i.e., a quarters ahead, we cannot assume that the outcome for the time points $t_{h+1}, \ldots, t_{h+a-1}$ are known at the time of prediction. Consequently, we have also excluded the corresponding $a-1$ objects, at time points $t_{h-a+1}, \ldots, t_{h-1}$, from the training set. Note that the latter only affects the second scenario in which we are making predictions four quarters ahead, i.e., resulting in that three preceding objects are excluded from the training set. Rather than searching for the optimal predictive model with careful hyperparameter tuning, we have here opted to use the standard *GradientBoostingRegressor* as implemented in [10], with default parameter settings.

Given the complete multi-variate time series $X \in \mathbb{R}^{N \times D}$, time points $T = (t_1, \ldots, t_N)$ and outcomes $O \in \mathbb{R}^N$, a model to make a prediction at the time point $t_h \in T_{b:N-a}$ for the outcome at time point t_{h+a} is generated from the objects $X_{b:h-a}$ and labels $Y = O_{b+a:h}$, where b is the number of lagged values and a is the number of time steps ahead for which a prediction is made. As stated above, we will consider $b = 5$ together with $a = 1$ and $a = 4$, respectively. Moreover, to handle the so-called *cold-start* problem, we will only make predictions for the last 64 time points for which the outcome a quarters ahead are known.

We will investigate the feature effect for a subset of the features over time. We have considered two options for aggregating the feature effect at time-point t_h; averaging the (absolute) feature effect using feature values in the entire training set, i.e., $X_{b:h-a}$, or averaging using feature values in a *time window* of size w, i.e., $X_{h-a-w+1:h-a}$. In this study, we will consider $w = 12$, i.e., values from the 12 most recent objects in the training set are used.

4 Experimental Results

In Fig. 1 and Fig. 2, the predicted vs. actual outcomes in million Swedish Krona (MSEK), are plotted when making predictions one and four quarters ahead, respectively, and when using the original (blue dashed lines) and transformed targets (orange dashed lines), where the latter concerns the percentage change, which here is projected back to the original scale to allow for a direct comparison. When predicting one quarter ahead, the model using transformed targets clearly outperforms the model using non-transformed targets, with the former obtaining a root mean-squared error (RMSE) of 20448 MSEK and a Pearson correlation coefficient of 0.981, while the latter has an RMSE of 37942 MSEK and a correlation coefficient of 0.936. The performance difference between the two methods is less clear when predicting four quarters ahead; using original targets leads to an RMSE of 36353 MSEK and a correlation coefficient of 0.954, while for the transformed targets the RMSE is 32918 MSEK with a correlation coefficient of 0.951. Interestingly, the task of predicting one year ahead is slightly more easy than predicting one quarter ahead when using the original targets. In terms of RMSE, the use of the transformation (percentage change) is clearly effective independently of the forecast horizon, while the correlation coefficient is hardly affected by this transformation when predicting a year ahead.

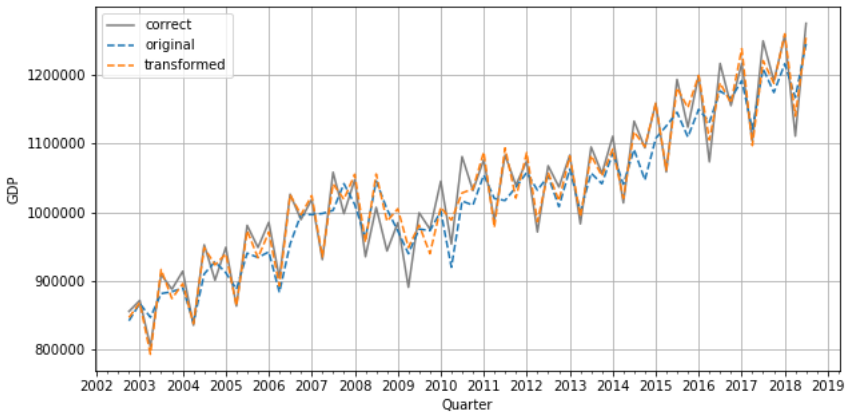

Fig. 1. Predicted vs. actual outcome one quarter ahead with transformed and original target

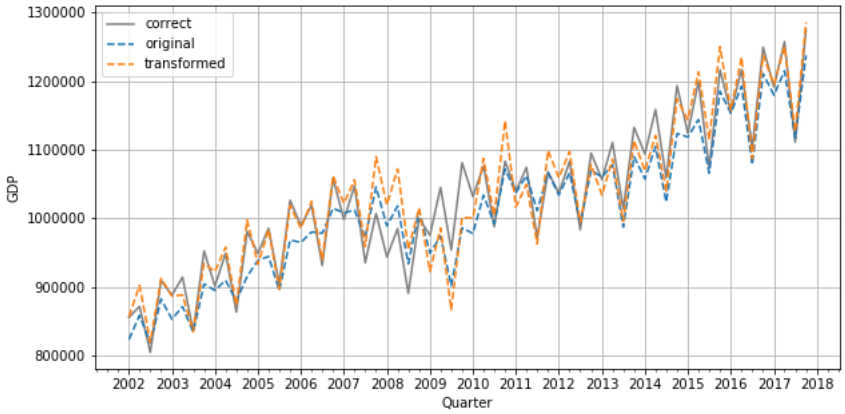

Fig. 2. Predicted vs. actual outcome four quarters ahead with transformed and original target

In order to analyze the differences between the models using the original and the transformed targets, we can take a look at the aggregated (absolute) feature effect, averaged over all predictions and for each prediction calculating the feature effect with respect to all previously observed feature values. In Fig. 3 and Fig. 4, the average absolute effect is plotted for the top 20 features (listed in descending order) when making predictions one quarter ahead, without and with the transformation, respectively. Note that although the scales differ, since the former model predicts the actual GDP (in MSEK) while the latter predicts the percentage change, we can still compare them by their relative impact. For the former, one of the lagged variables (`nbnpmpf-4`), which represents the outcome four quarters before the time of prediction, is dominating, while the effect is distributed differently for the latter, although four of the lagged variables appear among the top five.

Fig. 5 shows that when not having applied the transformation, the picture for predicting four quarters ahead is similar to when making predictions one quarter ahead; two of the lagged variables appear among the top five, although two different ones. However, when having applied the transformation, the picture changes quite drastically when predicting four quarters ahead, as shown by Fig. 6. Here, the highest ranked lagged variable (`nbnpmpf-4`) is ranked behind eight other features, which indicates that the importance of using lagged variables decreases when having applied the (percentage change) transformation and considering a more distant forecast horizon.

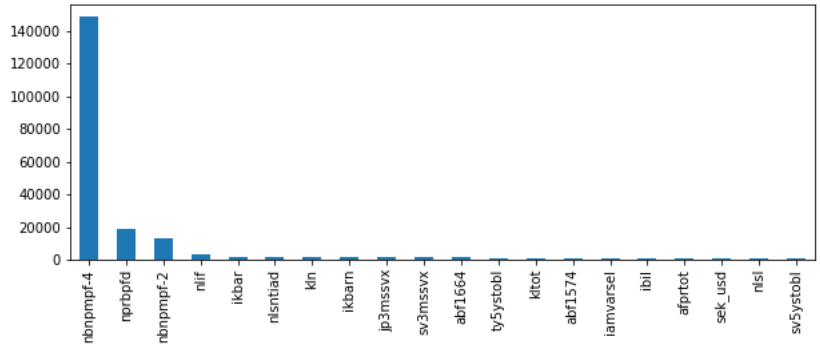

Fig. 3. Aggregated feature effect for original target and one quarter ahead

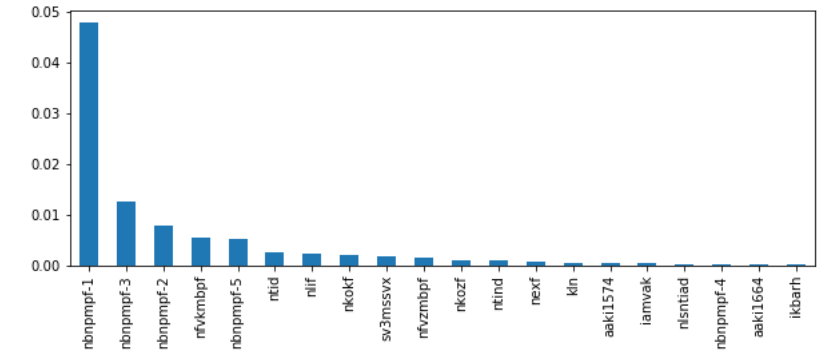

Fig. 4. Aggregated feature effect for transformed target and one quarter ahead

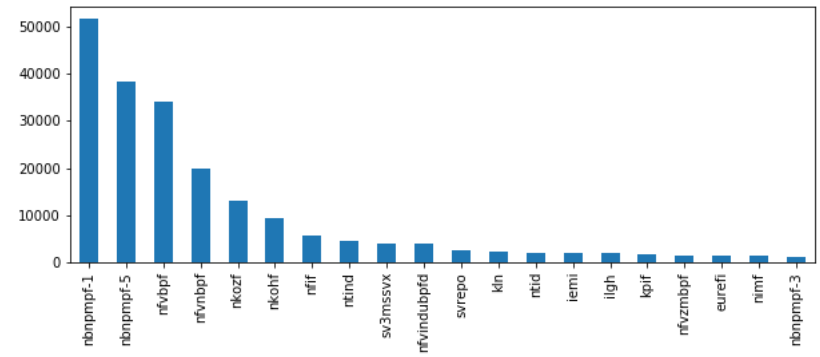

Fig. 5. Aggregated feature effect for original target and four quarters ahead

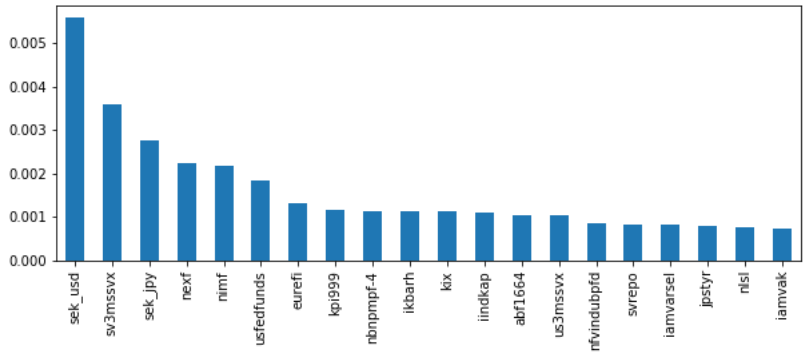

Fig. 6. Aggregated feature effect for transformed target and four quarters ahead

However, the graphs displaying aggregated feature effect do not show how it varies over time and can hence not be related to individual predictions. In Fig. 7, the feature effect for the nine top ranked features (according to Fig. 6), is displayed over time, where the feature effect is, as above, calculated with respect to all feature values that have been observed prior to each prediction. One may clearly see that the relative sizes of the effects vary over time.

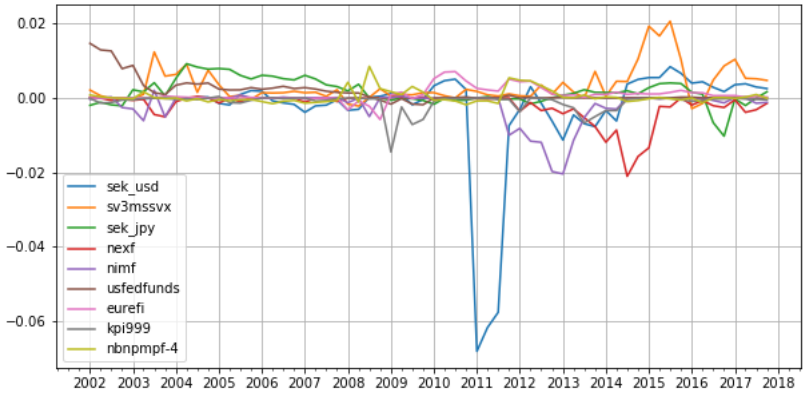

Fig. 7. Feature effect over time for transformed target and four quarters ahead

When calculating the feature effect in relation to all previously observed feature values, we do not know whether the impact is due to deviations to values observed a long time ago or more recently. One way of measuring feature effect in

relation to more recent observations is to include only the latest feature values in the calculation (limiting the set V in Eq. 1). In Fig. 8, the set of feature values to include is limited to the 12 most recent observations. When focusing the analysis to the most recent feature values, one may observe some dramatic changes to using all observed feature values, e.g., the effect of the Swedish Krona to US Dollar exchange rate (`sek_usd`) starts to increase significantly before 2010, when using the time constraint, while the impact of `nexf` (Swedish export) during 2014 disappears.

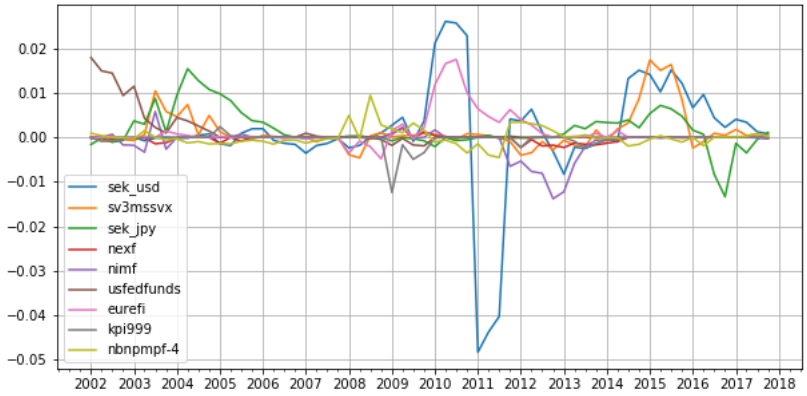

Fig. 8. Feature effect using time constraint for transformed target and four quarters ahead

When focusing on a specific prediction, we may take a look at an ICE plot (Eq. 2). In Fig. 9 and Fig. 10, ICE plots are shown for the predictions at January 1, 2010 and July 1, 2014, respectively, using the above features for which the values have been min-max-normalized to allow for displaying multiple features in one plot. The graphs show for both predictions that they are higher than the ones obtained with lower values of `sek_usd` and higher values of `eurefi` (Euro refinancing rate) and `sv3mssvx` (three-month rate in Sweden). The second prediction is lower than what is output by the model for all but high values values of `nexf` (Swedish export), while it is higher than what would be output for high values for the exchange rate of the Swedish Krona to Japanese yen (`sek_jpy`).

In an ICE plot, we can see how a prediction would be affected by replacing some specific feature value with all possible (previously observed) values. However, some feature values may be extreme and occur very infrequently. In addition, the values may also not have appeared for a long time, and hence may be of little relevance when reasoning about the current prediction. To allow for reasoning about the feature effect in a temporal context, we also take a look at the proposed ICE-T plots (Eq. 3), to study how the feature effect varies when

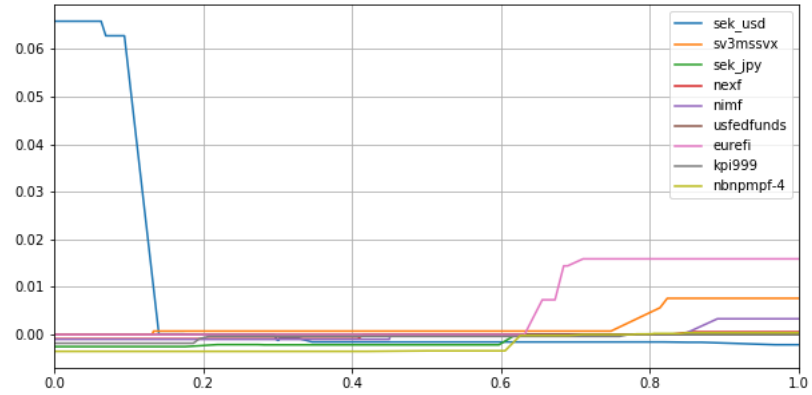

Fig. 9. ICE plot for 2010-01-01

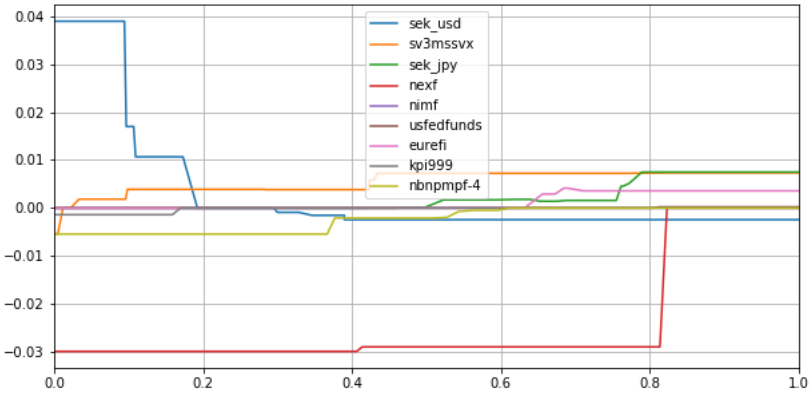

Fig. 10. ICE plot for 2014-07-01

selecting values in the order they have been observed. As a complement to the ICE plots, we present the corresponding ICE-T plots for the predictions at January 1, 2010 and July 1, 2014 in Fig. 11 and Fig. 12, respectively. For the first plot, we can see that the feature `sv3mssvx` (three-month rate in Sweden) only has an impact if considering feature values that occurred more than a decade before the time of prediction. Moreover, the effect of `sek_usd` is (very) high, only if considering relatively recent values, and this feature has hardly any effect if considering values observed earlier. In the second plot, one may observe that the prediction is lower than what is obtained when using values for `nexf` (Swedish export) that were observed more than 12 quarters earlier, which explains why the effect of this feature was not visible for the same date when calculating feature effect using the time constraint, as shown in Fig. 8.

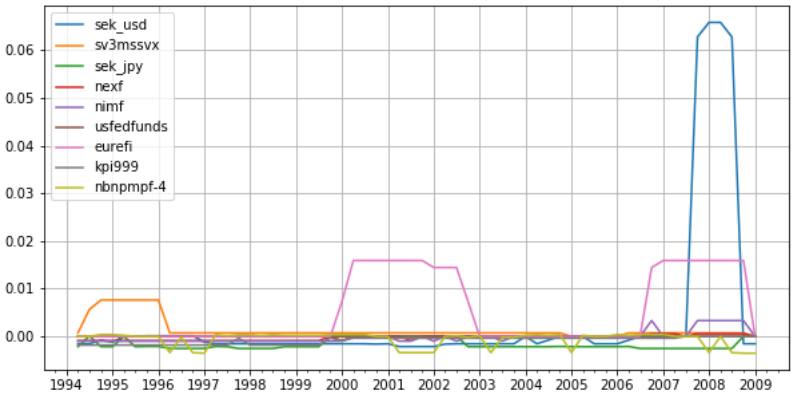

Fig. 11. ICE-T plot for 2010-01-01

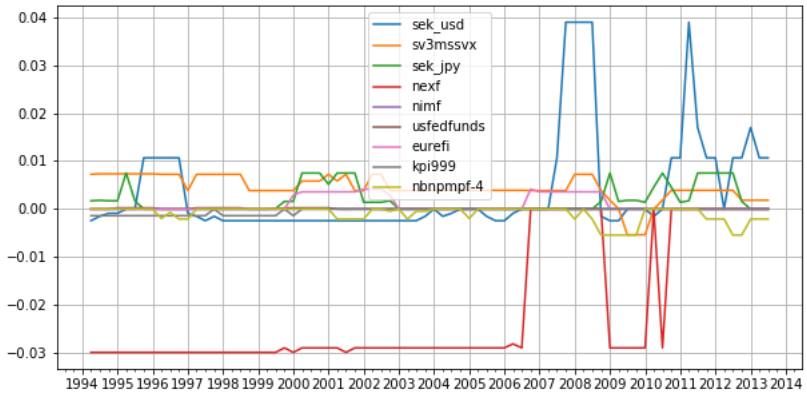

Fig. 12. ICE-T plot for 2014-07-01

5 Concluding Remarks

We have investigated ways to explain multivariate time-series forecasting models, by measuring and presenting the feature effect. In addition to calculating the aggregated feature effect, measured with or without a time constraint, we have presented an approach to visualize the feature effect on individual predictions, called the ICE-T plot, which complements the ICE plot by showing the feature effect over time. We have presented an application of these techniques to the task of predicting the Swedish GDP. In addition to demonstrating the use of the techniques, the empirical investigation has also highlighted the impact of target variable transformation and forecast horizon on the feature effect.

The work has focused on explaining predictions of a black-box model by analyzing the feature effect, i.e., how the output of the model changes when changing the input. For retrospective analysis, all the techniques considered in this work can be straightforwardly adapted to instead calculate feature importance, i.e., how the predictive performance, such as absolute error, is affected by changing the input. However, in a real prediction scenario, one does not have access to the correct target (for the prediction at hand), and hence the latter option is not available.

A natural extension of the current work is to consider the effect of changing multiple features simultaneously. In contrast to PDP and ICE plots, which need one axis per feature included in the combination, the ICE-T plots would remain two-dimensional, since the time points, rather than feature values, are used to align the predictions.

In the current study, we have only considered one specific type of black-box model (generated by gradient boosting), and a direction for future work is to study how the choice of black-box model (including various hyperparameter settings) affects what features are considered to have an impact. Finally, there

are several alternatives to using feature effect to explain predictions and the application and adaptation of these techniques to the specific requirements that multivariate time-series forecasting models provide could be a fruitful area of research.

Acknowledgments

The study was funded by Swedish Governmental Agency for Innovation Systems (grant no. 2019-02252). HB was partly funded also by the Swedish Foundation for Strategic Research (grant no. BD15-0006).

References

1. Economic tendency survey. Tech. rep., National Institute of Economic Research, `https://www.konj.se/english/publications/economic-tendency-survey.html`
2. European system of accounts - ESA 2010. Tech. rep., Eurostat, European Commission (2013)
3. Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001)
4. Chakraborty, K., Mehrotra, K., Mohan, C.K., Ranka, S.: Forecasting the behavior of multivariate time series using neural networks. Neural networks **5**(6), 961–970 (1992)
5. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Annals of statistics pp. 1189–1232 (2001)
6. Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E.: Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. Journal of Computational and Graphical Statistics **24**(1), 44–65 (2015)
7. Jung, J.K., Patnam, M., Ter-Martirosyan, A.: An Algorithmic Crystal Ball: Forecasts-based on Machine Learning. International Monetary Fund (2018)
8. Karlsson, I., Papapetrou, P., Boström, H.: Generalized random shapelet forests. Data mining and knowledge discovery **30**(5), 1053–1085 (2016)
9. Van der Laan, M.J., Polley, E.C., Hubbard, A.E.: Super learner. Statistical applications in genetics and molecular biology **6**(1) (2007)
10. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)