Last semester, I was working on a project to solve an age-old problem in economics. The Total Factor Productivity (henceforth, TFP) measures the rate of growth of a country. It is an extremely important metric in developmental economics, and is often used to compare how developed a country is with respect to others. Needless to say, the method of calculation of TFP is pivotal. Traditionally, TFP was calculated residually as proposed by the Solow model. What this meant is the following:

Country's output (Y) = Labour Growth Rate (L) + Capital Growth Rate (K) + TFP

In the above equation, every country has the values of Y,L,K and the value of TFP was merely obtained by subtracting L and K from Y. While this provided a neat way to calculate a country's TFP value, it had a major drawback. No information could be obtained on what the sources of TFP were, using this equation. Thus, given a TFP value, a country had no way of knowing in which sources to invest to bolster its TFP, and consequentially its output. Clearly, the relationship between TFP and its sources could prove invaluable to a country's development plan. Thus, the need for a model was felt.

The dataset I used for this project was a panel dataset of around 20 countries and 25-time years, therefore containing around 500 TFP values to be predicted.

Note: Most of the text is broken into points to be crisp and convey only the relevant information. If any of the lines have been taken directly from the paper or any other source, they have been ***italicized.***

Framing the question

Step 1: Finding a phenomenon and a question to ask about it

Goal: Model a relation between TFP and its sources

- Drawing from the clock analogy in the paper, sources are the gears of TFP. We are interested to figure what are the characteristics of the gears that make the TFP clock turn.
- The main goal of the project is to determine how is TFP affected by its sources and come up with a model for the same. We also want to determine what are the major sources which contribute to TFP.
- Focused model: It is important to focus on only the important sources, there might be many sources which marginally do improve a country's TFP but are not major contributors. Clearly, the government should not invest in such sources. Hence, a literature review to determine the major contributors is imperative and care should be taken to include only the relevant sources
- Avoiding fuzzy questions: It might be near impossible to come up with a generalized Sources-to-TFP model for the entire world. Clearly, for agrarian economies, the agricultural sector will be a major source of TFP and the

industrial sector might only be a menial contributor. On the other hand, for developed countries, industrial sector will be a major contributor whereas agricultural sector would be an almost insignificant contributor. Thus, the model will vary based on the type of country we are evaluating. The model in this project will focus on OECD countries with similar geographical, demographical and developmental features. A separate model will be required for emerging economies, another one for developing ("under-developed") economies and so on. Trying to find an umbrella model will lead to inaccurate results.

- Baseline: To improve over the Malmquist Index and other LP techniques (For economy of space, the details and drawbacks of these techniques are not mentioned here). Both techniques lead to complex mathematical models which lead to tough generalizability, and therefore none of these techniques are widely used over the traditional Solow model. The goal would therefore focus on generalizability, over exact matching of the predicted and actual values.
- Evaluation criteria – RMSE as we have defined a regression model

Step 2: understanding the state of the art

Gaps in literature

- Malmquist Index and other LP techniques – less generalizability, complex mathematical models
- To find sources of TFP, countries conduct qualitative analysis. This approach is, however, expensive and results depend on the availability of resources the country can afford to spend
- No previous machine learning approach to the problem

Step 3: determining the basic ingredients

- After a thorough literature review, where I analysed the different qualitative analyses of OECD countries, relevant sources were determined. For example, it was found the technological sector, human capital, exports and imports were major sources for TFP
- Sources were either time-varying – like suicide rates - or constant – like exports (stay almost constant over a large period of years)
- Missing values will be dealt with on the basis of whether the source is constant (replace with mean value) or time-varying (replace with last time-stamp value)
- Within transformation required on the dataset to eliminate cross-section collinearity

Step 4: formulating specific, mathematically defined

Hypothesis

- TFP depends on various economic sources (such as exports, human capital, suicide rates etc). Cross country effects of these are ignored to increase generalisability
- TFP = f(Sources of TFP), where cross-country effects are ignored, time-varying effects are considered
- The model should be built to obtain results that either prove or disprove the hypothesis, and the hypothesis should not be modified once model planning has begun to avoid HARKing.

Implementing the model

Step 5: selecting the toolkit

What modelling tools should be used and what level of abstraction is appropriate

- High level of abstraction: Do not need to know inner workings of FNN, loss or optimizer, a framework such as PyTorch or TensorFlow should be used

Step 6: planning the model

Different components: {FNN with inputs as sources and output as TFP, number of layers as 3, activation as LeakyReLU}, {loss function}, {optimiser}, {training, validation and test datasets}

Now each model box, icon, or flow can be considered individually, and its internal workings should be drafted in terms of mathematical equations. These should be explicit equations that can later be implemented in simulation programs.

Simplified equations for the model: hidden layer 1: h1 = W1*(Sources) + b1 →(X,15)

Activation 1: a1 = LeakyReLU(h1)

Hidden layer 2 : h2 = W2*(a) + b2 → (15,6)

Activation 2: a2 = LeakyReLU(h2)

Final Layer: y-hat = Wf*a2 + bf → (6,1)

Step 7: implementing the model

Importance of Unit Testing

- A major mistake that I made during my project was failure to unit test, which led to overfitting with no tangible cause (proper regularisation, and simple FNN)
- Unit testing should be done on each step, every layer, and using a simple optimiser like SGD
- Initially obtain results with no regularization, then add regularization such as Batch Normalization and Dropout and note the improvement in results in case no regularisation leads to overfitting. Starting with a regularised model can make it difficult to determine whether the model is generalised and whether the amount of regularisation is appropriate
- Simplest model: 2 hidden layer FNN with lr = 0.1, 50 epochs, tanh activation
- Later can be expanded to adaptive lr, batchnorm, dropout, more layers, or diff lr or epochs if required
- Plot model behaviour at each step – Loss plots should be done at every step

Model testing

Step 8: completing the model

- Stop once RMSE is around 0.5, this implies that our original hypothesis was answered and we were able to model a relationship between TFP and its sources. We also need to take into care that there is no overfitting, and ensure that our model is generalised by testing on the validation dataset.
- Don’t want complicated models as that leads to reduced explanatory power, need to be able to explain which weights affect more
- AIC/BIC criteria can be used to compare the neural network model with the typical LP model.

Step 9: testing and evaluating the model

- As a primary model, we do not require it to match the predicted values closely to the actual value, especially in the case of TFP a value of 90 is not significantly different from a value of 95.
- Thus, some amount of loss is acceptable.
- We have to ensure that there is no overfitting, by first evaluating the model on the validation dataset and then the test dataset.
- We expect the model to be generalizable as we removed cross-country effect using the within transformation.

Publishing

Step 10: publishing models

- Target audience: The paper can be published into top economic journals or in any of the several conferences on economic development. Since economic development is a relatively new field of research, researchers in the field are always open to new techniques and methodologies (For an estimate on how new the field is, Solow, the author of the "traditional" Solow model, died just 5 years ago :P)
- The FNN can be represented graphically for ease of understanding of the research community.
- All mathematical equations need to be provided, for the research community to adequately understand how the neural network is working to predict TFP values using the sources, and thus to provide credibility to the model used
- Both the data and the code should be publicly available to allow the researchers to pursue deeper understanding.