ACCESSIBLE AND REPRODUCIBLE RESEARCH

Data Science: from Academia to industry - supplementary report

Stephen Haben
Digital And Data Consultant (ESC)

Samuel Hinton
Senior Data Scientist (ARENKO)

JULY 2022

CONTENTS

DISCLAIMER

This document has been prepared by Energy Systems Catapult Limited. For full copyright, legal information and defined terms, please refer to the "Licence / Disclaimer" section at the back of this document.

All information is given in good faith based upon the latest information available to Energy Systems Catapult Limited. No warranty or representation is given concerning such information, which must not be taken as establishing any contractual or other commitment binding upon the Energy Systems Catapult Limited or any of its subsidiary or associated companies

1. INTRODUCTION

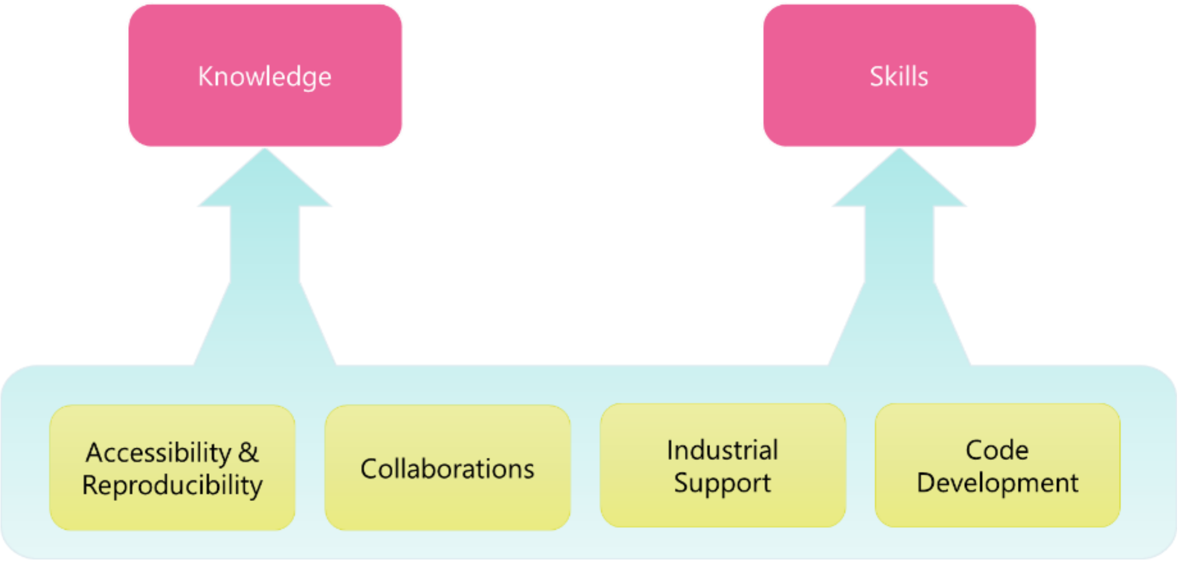

Figure 1. Illustration of the structure and the dependencies for our reports on academic driven impact within industry. The two main methods of impact are knowledge and skills (in pink) and these are supported by four supportive mechanisms (yellow). Each of the supportive mechanisms is explored in more detail in supplementary reports. This report focuses on Accessibility & Reproducibility.

The research produced in academia can have positive and wide ranging impacts industry and is essential for supporting the innovative data science products and services if the energy sector is going to achieve Net Zero. The main report discuss the two main impacts that academia can have in industry:

- Technical knowledge, and
- Skills

This report is one of four supplementary texts which delve into supporting mechanisms for these types of impact. This report considers **Accessible and Reproducible research.**

For the knowledge generated via academic research to make real impact in industry it should not only be openly available, but should provide enough information so that the methods and approaches to be accurately recreated. As discussed in the main report (Haben, &, Hinton, 2022b), many of the industrial participants we interviewed struggled with accessing research because either it wasn't very visible, or perhaps sat behind a paywall. There are new emerging models which focus on open research principles (i.e., openness throughout the research process including data, code, methods etc.). Platforms such as F1000 focus on open access and open data but also on opening the peer-review process in contrast to the usual blind reviews currently common in academia.

Even when companies could access research, or at a preprint, many times the manuscripts were written with inaccessible language, or the methods were simply not clear enough to replicate the results. This report dives more deeply into these issues, opportunities and challenges, and provides some guidance on what makes a good quality, reproducible research paper when describing methods and techniques. In addition to the responses from interviewees this section also shares some insights from a literature review we performed to identify common mistakes, shape the guidance, and identify exemplars. The investigation of the literature has focused on price forecasting since this is one of the primary elements of the authors experience and strongly aligns with the work performed by Arenko.

Note, there is a lot of other and comprehensive guidance available on proper procedures for reproducible research in general, in particular the "Turing Way" (The Turing Way, 2022), an open source, collaborative and community driven project which interviewees identified as a useful resource. It also shares best practices in coding, data management and version control. This report focuses on particular issues in the area of data science within the energy systems sector.

2. ACCESSIBLE AND REPRODUCIBLE RESEARCH

Section 4.4 in the main report (Haben, & Hinton, 2022a) outlined the issues with finding and accessing relevant research for the problems industries are trying to solve. In this report we consider accessibility to mean the issue with trying to access the knowledge and information within the research. This is a prerequisite to reproducing that research and hence the ideas are very closely aligned. For research to be utilised by academics the methods should be reproducible in a reliable way. Since many algorithms have a stochastic component, an exact reproduction may not be possible but qualitative features such as relative accuracy should remain if the methodologies and tests are appropriate and reliable. Well-written methods also increase the trustworthiness of outputs.

The aims of this supplementary report are to highlight what our literature review and interviews identified as clear blockers to being able to understand and recreate the models within their own industry. Without reproducibility then the research is largely redundant and hinders further research and investigation since extension and upgrades to the work are not possible.

We consider the following points:

- Data: real data is key to proper data science. Models cannot be trained, tested or validated without access to good quality open data. This will be considered in Section 2.1.
- Papers: The primary description of an academically derived method or techniques, is via the published manuscripts. What information is included and how it is described is required to enable uptake or honest recreation of the models. These are discussed in Section 2.2.
- Code: Although releasing code is not a priority for academics. The sharing of a code (or even pseudo-code) can be very beneficial to reproduction of a model. This is the subject of Section 2.3.

Open data and modelling in energy research is vital but appears to need improvement (Pfenninger, DeCarolis, Hirth, Quoilin, & Staffell, 2017). The next few sections identify some of the challenges and potential solutions in producing reproducible results in the main areas as listed above: data, papers and coding. Tools which can support accessible and reproducible research are presented in a supplementary report on coding development for academics (Haben, &, Hinton, 2022c).

2.1. DATA

Data is a vital component to produce research outputs in the first place. Without it models cannot be properly trained or designed, and they are unlikely to be accurate or useful for the applications intended. Data availability is a big problem in data science research. Much of the data comes from innovation projects which means that they are typically short in duration, may been modified by interventions (and hence do not represent the true condition of the real data), and they are often only allowed to be used with the project team. In other words, if you are not part of the project, then you are unlikely to be able to use the data. This can be seen as beneficial for those working on the original project. They would like first pass on the data with the possibility of being the first to publish work based on it. One solution to this is to have a delay on releasing the data used on a (especially publicly funded) project.

As an example, a recent review of load forecasting methods for low voltage level network demand (Haben, Arora, Giasemidis, Voss, & Greetham, 2021) found that less than 24% of the papers present results using open data, and of those 42% used data from a single project. This not only means that it is impossible to recreate the results from most papers, but the other papers use a limited number of data sets (likely because there aren't many out there) which means they haven't been tested for how generalisable the results are. In other words, the methods are not proven to work.

Open data is dependent on the data providers. This is one way in which industry can better support academia and the research that they perform (See supplementary report on Industrial support for academia (Haben, & Hinton, 2022d). Since the publication of the Energy Data Taskforce (Sandys, et al., 2019) and the recommendation of Presumed Open, the movement towards opening and sharing more data, especially from the energy networks, is rapidly increasing. More is still required however, especially of smart meter data, which have been shown to be very valuable for smart grid research. Many journals are making open data a required element of the publication process[1]. Although this may encourage the increased use of open data sources, it should not come at the sacrifice of learnings from data if they cannot be published. Sharing data can also increase the citations associated with a paper which should also encouraging to the academic community (Colavizza, Hrynaszkiewicz, Staden, Whitaker, & McGillivray, 2020).

There are other properties of data which are valuable for reproducible research. In particular, **Data should:**

- **Be linked** with the original corresponding publication. This allows a user to recreate the whole process, from cleaning, to methods, to testing.
- **Be modern**. Ideally it should be within months of the publication of the paper, not multiple years. As new technologies and energy efficiency products are implemented energy data can quickly go out of date and be less relevant to the situations and applications for which the methods are being developed.
- **Ensure realism**. This was a common issue in papers that may not be aware of the real-time constraints of data availability. For a simple example, some papers reviewed attempted to predict UK imbalance price outturn, and found the best predictor was the imbalance price from the previous settlement period. This is correct, however from an industrial application, this is a fatal issue. Imbalance prices are not available for an hour after the settlement period, so a real-time model would be unable to access the prior imbalance price. Subtleties about delayed data publication are often lost in historical datasets, especially when they are reduced to tabular data.
- **Be open.** For reproducibility, data should be sourced if possible from open datasets that as many people as possible can access. This has a drawback in that often other supporting data is removed to ensure anonymisation is enabled. If a closed data set is used, an equivalent open data set should also be tested with the model. This provides more information on generalisability but also allows more replicability and conclusions could be cautiously extended to the closed results.
- **Be provided.** If licensing permits, data should be archived on a data platform to allow for simple or automated download. If not, instructions should be included on how readers can attain the same dataset. For example, rather than just provided the name of the provider, authors can provide a link to the download page, and stipulate what parameters they passed to the data provider.
- **Be diverse.** Testing on a single data set is limited to prove the generalisability of your data set. Further there may be different effects depending on customs, weather, climate, socio-economics, etc. Testing across a diverse and representative sample of datasets ensures

[1] https://arstechnica.com/science/2021/11/keeping-science-reproducible-in-a-world-of-custom-code-and-data/

confidence in the models and results, but also identifies limitations and what may cause them.

- **Have sufficient size.** There were several papers reviewed wherein complicated techniques like deep learning were applied on datasets of dozens or hundreds of rows. Data should be of a sufficiently large time range to properly train the models being investigated. This could depend on the application and other parameters in the model.
- **Be linked to** other datasets where possible. For example, weather data is a potentially strong driver of demand and price. Although in some cases you may be able to source the weather yourself, in some instances you may not be able to. For example, the location data is removed due to anonymisation or pseudonymisation. In which case it would be useful if the data providers could link to important covariates or dependencies. They should also include any calibrations that they used.

Storing the data can be problematic especially with the increasing size. However, there are options available even for datasets up to terabytes in size. Zenodo, run by CERN, promises to store the data "safely for the future in CERN's Data Centre for as long as CERN exists." Further each data source uploaded is assigned a Digital Object Identifier (DOI), allowing it to be easily shared and cited.

2.2. PAPERS

The interviewees note that there is a large amount of poorly written journal papers, and in addition many papers are inaccessible to non-experts due to being written in very technical language and/or in an abstract way. A major reason for this is that the review process (and reviewers who themselves often have very little industrial knowledge) focuses on novel research and complexity and will reject many papers despite their practical and usefulness from an industry viewpoint. A common complaint was the lack of proper benchmarking and unclear methodology. Often it is impossible for a reader to recreate the methods presented and therefore there is a loss of trust in the methods.

Without a proper benchmark, it is not easy to understand the true value and quality or accuracy of a model. Good benchmarks should be easy to recreate but also be reasonably competitive which allows other papers to compare against a common model but also helps to understand the state-of-the-art and possible ways of improving the models. Such benchmarks would be very valuable in ensuring that new innovators or new teams can have a head start in developing their models and approaches. Further sharing the code (see Section 2.3) would also help practitioners.

There were also several other issues interviewees from industry had with academics papers, including:

- **Be as realistic as possible:** There is very little useful information in a paper which tackles a completely unrealistic scenario or constraints. If an academic author wants to make industrial impact then they must listen (or better collaborate (Haben, & Hinton, 2022b)) with the industrial experts who will use their research. This means including the framing and constraints that are most important to them and not necessarily the ones which make the research more convenient.
- **Being too long:** some journals have page limits while others can allow larger numbers of spurious and/or technical details.
- **Computational power to run:** for the practitioner this can be deal-breaker from the start if it is computationally intensive. If a complex model only marginally outperforms a simple

benchmark, but incurs far more technical debt and computational cost to run, often the benchmark is the more viable model.

- **Using inappropriate error metrics:** Without properly chosen metrics the results may have little to no meaning for the application of interest.
- **Using old data:** With changing technologies such as electric vehicles and distributed generation, not to mention the churn of business and households, energy demand can rapidly change. Hence results can quickly become invalid. Using up to date data, if available, is always preferable (Section 2.1).

As part of this research, we reviewed a number of papers to better understand some of the issues and problems with producing accessible and reproducible papers. A Scopus query was used to select highly cited papers in the areas of electricity price forecasting and predicting, and the most impactful **non-review** papers since 2014 were investigated. These papers were reviewed by a team of industry data scientists from Arenko with the goal of quantifying how value could be extracted from each paper and put into industrial practise. More complete details of the process are included in Appendix 4.

Even among highly cited papers, there were significant problems that commonly occurred in a the less cited ones. In none of the papers reviewed was the data provided directly, or available as an archive on a data platform like Zenodo, and the data used was commonly multiple years old at the time of publication. No papers provided code with documentation, nor implemented coding standards. Only one provided any code at all.

Without the data and code to allow reproducible results, the review focus turned to ascribing value to the methods in the reviewed papers. For example, a paper might claim a highly accurate machine learning prediction, but whether this is due to the data range, auxiliary variables, data processing methodologies, or the machine learning or statistical model is a key piece of information for those looking to extract the value from the paper. Only a single paper reviewed attempted to break down the metric improvements from different components of the analysis of a whole.

This finding meshes with the interviewee responses, in that many published papers chase metrics, and present cherry-picked data ranges with exotic hybrid models that happen to beat an unimpressive benchmark over a short time range. In total, we identified numerous areas which would dramatically increase a paper's academic and industrial usefulness.

Feature engineering should:

- **Explicitly state data splits.** Papers should explicitly detail the split into training, test, and validation datasets.
- **Use realistic data available at the time.** The training and testing datasets should have no overlap and should also incorporate realistic gaps. For example, if a researcher is utilising a model to predict wholesale prices, there is usually a specific gate closure (e.g. midday the day before). In this case the training data should not include the 12 hours prior to the test set to ensure the accuracy of the forecast does not use any data which is unlikely to be available. This is the same point as made in Section 3.1, "Ensure realism". If predictions need to be made for a given time, do not include data available only after that point in time.
- **Have their importance's made explicit.** Models which include multiple features should have feature importance explicitly stated or at the very least investigated. There are many

established methods to investigate feature importance, like Shapley values, however only a minority of reviewed papers tried to do this.

Modelling should:

- **Have a baseline/benchmark.** Roughly half the papers reviewed did not include a simple baseline to compare model performance against.
- **If possible, include a state-of-the-art comparison.** Too often, papers compared their machine learning model performance against simple statistical models like autoregression or linear regression. Ensuring that models compare against state-of-the-art competitors allows is far more useful, even if more difficult to show positive results. This point has been raised by many interviewees and prior reviews such as (Lago, Marcjasz, Schutter, & Weron, 2021) who even provide code implementations for state-of-the-art models to try and drive adoption.
- **Use proper metrics.** Many papers borrow metrics from historically highly cited papers, even if there are better alternatives available. In the more modern reviewed papers, the majority of papers followed this, and only a small number used inappropriate metrics like MAPE for prices (prices often cross zero which means that the MAPE is not defined).
- **Use proper cross validation.** In time series models, rolling window cross validation (also known as time series cross validation) provides the best insight into model performance in industrial applications, where models may be retrained automatically or at a regular frequency. The majority of papers reviewed used the simpler single train-test-validation split rather than rolling window cross validation. Additionally, a significant number of papers reviewed indicated they made use of scikit-learn, and the more advanced rolling-window cross validation functionality in TimeSeriesSplit has been available since 2016. It is unknown whether many authors were unaware of the need to perform this validation, unaware that it has been implemented by a trusted third-party library, or decided it was not worth the extra complexity.
- **Detail their hyper parameter optimisation.** Feeding into the original point about trying to determine wherein the paper the value lies, models should have their performance during hyper-parameter optimisation clearly listed to both justify the final model architecture and show the generalisability of the method. Ideally the hyper parameters themselves should be shared so as to enable recreation of the results. This could be included by sharing the saved model (Section 2.3).
- **Do significance testing.** The majority of reviewed papers concluded with a textual comparison of metrics. With multiple metric comparisons, papers should employ statistical tests such as the Diebold-Mariano test (when applicable) to show how meaningful their results are.
- **Focus beyond point forecasts.** Papers aiming to have the highest impact should move beyond point forecasting into more sophisticated probabilistic forecasting. Uncertainty on model predictions is often overlooked in papers, but is a priority in industrial applications.

As also discussed in Section 4.5 in the main report (Haben, & Hinton, 2022a) an important way to ensure that the paper is reproducible is making it open. The inaccessibility of many journals, especially to industry partners, has been addressed previously. We include it here as well for completeness.

For two examples of papers that at the top of our reviewed list, see the review paper from (Lago, Marcjasz, Schutter, & Weron, 2021) who provide both data and code out of the box for future papers to use, and (Ugurlu, Oksuz, & Tas, 2018) who apply recurrent neural networks onto price forecasting, and touch on hyperparameter tuning, feature importance, rolling window validation, and interactive model development to make the valuable sections explicit. Some of their important features are described in our case study in Appendix 4.3.

It is important to remember that reading academic papers also requires experience and skill and therefore, as one interviewee mentioned, can be inaccessible to many within their organisation. An easy way to make academic writing more accessible is to produce more accessible versions of the article, e.g. through blog posts, or social media breakdowns, or generating a conference proceeding which focuses on the practical aspects. It may be also worthwhile for publishers to introduce journals with a more practical focus.

2.3. CODE

Code can be extremely useful for helping to decipher or elaborate on the methods described in the paper. Reproducibility is a major issue. For example, in the popular M-time series forecasting competitions, for the third challenge, "although the test data and the submitted forecasts are all available publicly, the computed accuracy scores do not match those in the published paper" (Hyndman, 2020). For this reason, the next competition, M4, required participants to also submit their code in a Git repository. Analysis from other less related fields have also found similar problems. For example, even for relatively simple statistical analysis, an analysis of over 200 papers found that their calculations could not be replicated with the information made available (Weissgerber, Valencia, Garovic, Milic, & Winham, 2018).

Journals have increasingly started to request code to be included in the peer review process, especially where bespoke programmes are being applied. Further, many providing guidelines to help support those wanting to share code[2]. There are also initiatives like "Papers with Code", an open-source resource for machine learning papers. As mentioned in section 4.5 in the main report (Haben, &, Hinton, 2022a), there is very little incentive for an academic to share their code and hence more recognition, citation indexes would encourage further uptake and highlight the impact that the work is producing.

Other benefits to sharing the code and allowing others to add edits and raise issues is that it can drive continual improvement and testing the code without much further work. It will also increase the impact of your work, and potentially help you be hired since it is an open demonstration of your abilities (Section 5 in the main report (Haben, &, Hinton, 2022a)).

In the review that we conducted to demonstrate and better understand reproducibility we found very few examples of shared code, including no examples on "Papers with Code" for load or price forecasting. However, the price forecasting review paper (Lago, Marcjasz, Schutter, & Weron, 2021) does include an associated open-access price forecasting package, with associated benchmarks, datasets and metrics. The package comes with a GNU Affero General Public License, which means any modifications "used to provide a service over a network, the complete source code of the modified version must be made available". This is very beneficial for the community of researchers

[2] E.g. the high profile journal Nature has the following guidelines https://www.nature.com/documents/GuidelinesCodePublication.pdf

who can build upon the work and share their working but may not be as desirable for industry who obviously may want to reveal the updates and upgrades they made for their organisation.

Rather than sharing the code, another desirable alternative mentioned by interviewees was the possibility of sharing trained models. This is common practice within the computer vision and especially the Natural Language Processing (NLP) community which has the Hugging Face platform which enable trained models to be shared and hence repeated by anyone.

If code or models are shared, then ideally good standards of practice need to be abided to. Code skills and training will be covered in another supplementary report (Haben, &, Hinton, 2022c), but professional code can not only give confidence in reliability of the modelling, but it can aid reproducibility as a user can clearly follow the steps and reasoning behind each implementation. It is unlikely that many industrial organisations will blindly apply academic code and many want to reimplement the program if they are to used for business purposes.

In summary, from our interviews and the literature review, the following is some of the importance criteria for codes developed as part of research. **Code should:**

- **Be available.** Whether it's provided on a version control system (VCS) provider like GitHub or provided in a totally reproducible binder, code is authoritative and the foundation of almost all ML papers. Code not being provided reduces trust in the work and being unable to validate the results directly often means surprising outputs can result in the code being dismissed.
- **Have some documentation.** Code which provides no instructions on how to run, what the outputs are, and how complicated pieces of logic function, can have little value.
- **Follow some standards.** Whilst expecting all code repositories to adhere to best software engineering practises would be unreasonable, code should have some minimal structure and standards applied to make its digestion easier. Tools like Black exist which can make tasks like code formatting require no manual work and increase the readability of code.
- **Include dependencies.** At the bare minimum it would be useful to include versions of each package that the developers used when running their code but including specifications of their machine would also be useful for understanding performance, practicality and validate any reproduction.

There are many tools which can help assist with an academic wishing to open their code. Some of them were discussed in our interviews and are listed in the third supplementary report on coding development for academics (Haben, &, Hinton 2022c).

3. SUMMARY & RECOMMENDATIONS

Companies use academic outputs for a variety of purposes, this could be to help solve a problem which they are currently working on, to generate new models for their applications, or to create entirely new products for their business. In addition to the inaccessibility of academic research due to paywalls and lack of visibility, academic outputs may not be accessible due to the way they are written. This includes the complexity of their contents (there can be sometimes a certain pride in making language and methodologies more complicated than necessary) or they may be extremely lacking in the details such as how they have trained, selected, or applied their models or techniques. This means to understand the contents, it may require employees who are particular erudite in the topic or some blind extrapolation of the material.

As part of this topic in this report we also conducted a short literature review of papers in the area of price forecasting, from this and the interviews we wrote some guidance on how papers and outputs should be written and what they should contain. Some of the best practices include:

- **Data availability:** Papers should be released with the data or should use open data which should be link to the manuscript. It is vital to help reproduce and test the results and helps others to develop the methods further. Improvements in making data more open is a major way this can be supported.
- **Code sharing:** Can be a game changer in revealing the underlying method and providing key insights which are not apparent in the paper. This can be particularly useful if the paper is largely theoretical, and the implementation details are not included. Although it is rare for an organisation to use a code directly within their business it is usually a good starting point for them to develop their own code or insights.
- **Accessible, assessable, and reproducible papers:** As well as being open access, papers should be written so that all details of the methodology are clear. A simple rule of thumb is that a method should be reproducible from the contents alone. This may not always be possible due to any methods which include some degree of randomness, but the results should be robust enough that this does not affect the overall outcomes.

4. APPENDIX: OVERVIEW OF PAPER REVIEW

This section describes some of the details of the paper review performed by the Arenko data science team for this project. The focus was on electricity price forecasting is an area of expertise and importance for the team. Below we describe the search which was performed, and some of the criteria used to access the quality of the papers. We finish this appendix with a case study looking at some of the features of two papers from the review which illustrate the criteria and are exemplars of better research we have found. A detailed review is not presented here as the focus is on the general benchmarking criteria which make a quality paper rather than individual scrutiny of each paper.

4.1. SCOPUS SEARCH

To select the papers for the review, we use SCOPUS, Elsevier's abstract and citation database. This is a common database used for review papers. SCOPUS has a search option which can take various combinations of search words to filter the initial papers. Our initial search was as follows:

```
TITLE ((“forecast*”  OR  “predict*”)  AND  “electric*”)  AND  TITLE-ABS-
KEY ( “day ahead”  OR  “day-
ahead”  OR  “spot”  OR  “elexon”  OR  “balancing”  OR  “imbalance”  OR  “
nord”  OR  “price forecast*”  OR  “forecast* price”
)  AND  PUBYEAR  >  2014
```

This search is similar to those used in other reviews (Haben, Arora, Giasemidis, Voss, &, Greetham, 2021) and ensures that the widest selection can be selected and the focus is on the most recent papers. Additional manual filtering can then be applied where the papers are not relevant.

4.2. SOME CRITERIA FOR QUALITY PAPERS

Below is a list of criteria of what makes a quality paper which can be understood and replicated by users. This list was populated from both *a priori* knowledge from the authors and reviewers, and from further features which became apparent during the review process. They consider finite categories including: Data; Features; Modelling; and Code. The table shows the Category, Name of the criterion and a short description of the feature.

Category	Name	Description
Data	Modern	Ideally data is from the last few years. Data might be modern to when paper was published, but not modern anymore.
Data	No peaking	Data is available at time of predictions, respecting availability, trading times, data publication time etc.
Data	Open	Data is publicly available.
Data	Provided	Downloadable on Zenodo or other data platforms. I.e. you don't have to figure out how to get the data yourself, they make it explicit
Data	Size	Sufficient size for intended purposes, including for training and testing, and to take not account seasonalities and other features in the data.
Features	Split explicit	Train, test, validation split is explicitly stated in the paper.
Features	Realistic data split gaps	Realistic gaps and no overlap between training and testing data.
Features	Importance	Feature importance explicitly shown, including importance of lagged components if included.

Modelling	Assumptions	Free from overly simplistic assumptions about data/model/features.
Modelling	Baseline	Compared to a sensible baseline/benchmark. Baselines should be investigated.
Modelling	Simple comp	Compared to simple statistical models like AR
Modelling	SOTA comp	Compared to state-of-the-art models (LEAR, DNN, etc)
Modelling	Metrics	Compared using sensible/multiple metrics appropriate for the application/data
Modelling	Rolling Window	Models compared using proper rolling window cross-validation and not just a naïve "train-validation-test" split
Modelling	HPO	Is HPO done and documented properly. Should at least be performed and ideally difference between trials should be detailed somewhere.
Modelling	Significance	Metric comparison checked for statistical significance test, e.g. Diebold-Mariano test.
Modelling	Uncertainty	Model uncertainty investigated and quantified.
Modelling	Probabilistic	Models focus on probabilistic results rather than point forecasts
Modelling	Reliability	Uncertainty/probabilistic analysis investigating both reliability and sharpness, e.g. through proper scoring functions.
Modelling	Iterative	Benefits from data, cleaning, feature extraction, model, HPO, all clearly broken down.
Code	Available	Code hosted on public sharing repository (GitHub etc)
Code	Documentation	Code documented with readme, docstring, etc
Code	Standards	Code adheres to common SE standards (DRY, SRP, modularity, formatting, etc)
Paper	Open	Paper available without paywall in some fashion (open-source paper, or preprint available through arxiv, etc)

4.3. CASE STUDY ON EXEMPLAR PAPERS

As part of the review two papers were found to demonstrate many of the principles which we found adhered to many of the criteria described in Section 4.2.

4.3.1. FORECASTING SPOT ELECTRICITY PRICES: DEEP LEARNING APPROACHES AND EMPIRICAL COMPARISON OF TRADITIONAL ALGORITHMS (LAGO, RIDDER, & SCHUTTER, 2018)

This paper provides a comparison between different statistical techniques and machine learning algorithms and their application to forecasting electricity spot prices. The paper includes clear definitions and disambiguation of acronyms, diagrams to help explain common networks, and provides rigorous statistical checks on model performance beyond comparing a specific metric.

The data utilised in the model is open access, the split between training and testing is explicit, and appropriate for timeseries data. Models are compared not only to each other, but to a naïve baseline, and compared using multiple appropriate metrics. On top of this, many of the models investigated are commonly parametrised, and some basic hyperparameter tuning has been explored, with optimal hyperparameter results shown clearly in tabular format.

For anyone looking into machine learning and its application to the energy market, this paper represents a fantastic introduction. It could be improved if the code used to perform the benchmarking and hyperparameter tuning was made open-source. Finally, industrial application of machine learning model in forecasting has a strong focus on uncertainty and probabilistic predictions, however all models investigated in the paper provided point forecasts.

4.3.2. ELECTRICITY PRICE FORECASTING USING RECURRENT NEURAL NETWORKS (UGURLU, OKSUZ, & TAS, 2018)

Ugurlu, Oksuz, & Tas, provide an in-depth look at the applicability of various temporal neural networks to the Turkish day-ahead electricity market. The data split between training and test is explicit, and–unlike many similar papers–uses rolling window cross validation, which is a more appropriate choice than simply creating a global train and test split on the dataset.

The importance of various features passed into the models is explicitly evaluated. These features (lagged price values) are explored in an iterative approach across multiple models, where tabular data is shown with the four evaluated models (CNN, ANN, LSTM, GRU) and their performance metric (MAE) when subsequent features are included. This exploration into feature importance is rarely seen in academic publications, however, offers valuable insight for the readers of the article if they wish to implement the findings of the paper. All too often, promising results present a single, monolithic entity comprised of the data, engineered features, model choice, and hyperparameter selection. Adding iterative details like feature importance allows readers to extract the valuable findings from the paper without having to guess which of the aforementioned steps provides significant value.

Additionally, the models are compared to a baseline and simple statical models like SARIMA. The machine learning models were also compared to each other using the Diebold-Mariano test (as was also the case with (Lago, Ridder, & Schutter, 2018) to check for statistically significant improvement). As with (Lago, Ridder, & Schutter, 2018), this paper could be improved if the code behind it was made open source, and potentially extended into probabilistic forecasting.

5. BIBLIOGRAPHY

The Turing Way. (2022). Retrieved from The Turing Way: https://the-turing-way.netlify.app/welcome.html

Colavizza, G., Hrynaszkiewicz, I., Staden, I., Whitaker, K., & McGillivray, B. (2020). The citation advantage of linking publications to research data. *PLOS ONE*.

DuBois, J. (2020). *The Data Scientist Shortage in 2020.* Retrieved February 8, 2022, from https://quanthub.com/data-scientist-shortage-2020/

Elsevier. (2022). *Scopus*. Retrieved from https://www.scopus.com/home.uri

Haben, S., Arora, S., Giasemidis, G., Voss, M., & Greetham, D. (2021). Review of low voltage load forecasting: Methods, applications, and recommendations. *Applied Energy*.

Haben, S., & Hinton S. (2022a). Data Science: From Academia to Industry – Making Impact in the Energy Sector, Energy Systems Catapult

Haben, S., & Hinton S. (2022b). Academic and Industrial Collaborations, Data Science: From Academia to Industry – Supplementary Report. Energy Systems Catapult

Haben, S., & Hinton S. (2022c). Code Development for Academics Entering Industry, Data Science: From Academia to Industry – Supplementary Report. Energy Systems Catapult

Haben, S., & Hinton S. (2022d). Industrial Support for Academics, Data Science: From Academia to Industry – Supplementary Report. Energy Systems Catapult

Hyndman, ,. R. (2020). A brief history of forecasting competitions. *International Journal of Forecasting*.

Lago, J., Marcjasz, G., Schutter, B., & Weron, R. (2021). Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark. *Applied Energy*.

Lago, J., Ridder, F., & Schutter, B. (2018). Forecasting spot electricity prices: Deep learning approaches and empirical comparison of traditional algorithms. *Applied Energy*, 386-405.

Pfenninger, S., DeCarolis, J., Hirth, L., Quoilin, S., & Staffell, I. (2017). The importance of open data and software: Is energy research lagging behind? *Energy Policy*, 211-215.

Sandys, L., Dobson, R., Brown, E., Graham, G., Lane, R., & Verma, J. (2019). *Energy Data Taskforce: A Strategy for a Modern Digitalised Energy System.* Energy Systems Catapult.

Sandys, L., Dobson, R., Verma, J., Johnston, G., Roberts, D., Leland, B., Haben, S., Ainsworth, E., Guinta, F., & Pearson, S. (2022). *Delivering a Digitalised Energy System: Energy Digitalisation Taskforce Report.* Energy Systems Catapult.

Ugurlu, U., Oksuz, I., & Tas, O. (2018). Electricity Price Forecasting Using Recurrent Neural Networks. *Energies*.

Weissgerber, T., Valencia, O., Garovic, V., Milic, N., & Winham, S. (2018). Meta-Research: Why we need to report more than 'Data were Analyzed by t-tests or ANOVA. *eLife Sciences*.

OUR MISSION

TO UNLEASH INNOVATION
AND OPEN NEW MARKETS
TO CAPTURE THE CLEAN
GROWTH OPPORTUNITY.

ENERGY SYSTEMS CATAPULT
7TH FLOOR, CANNON HOUSE,
18 PRIORY QUEENSWAY,
BIRMINGHAM, B4 6BS.

ES.CATAPULT.ORG.UK
@ENERGYSYSCAT