NeuralProphet project for FDS course

Polina Pilyugina [1] **Alexey Voskoboinikov** [1]

Abstract

This project aims to contribute to the open-source library NeuralProphet by adding state-of-the-art models and refactoring the code to adopt the best machine learning engineering practices.

1. Problem Statement

NeuralProphet is a new library for time series forecasting built on PyTorch. It is inspired by the widely known Facebook library for time series forecasting called Prophet ((Taylor & Letham, 2018)) and DeepAR model ((Salinas et al., 2020)). However, while Prophet is an additive model focused chiefly on seasonal components and holiday effects, NeuralProphet additionally includes AutoRegression components. Moreover, NeuralProphet is built on PyTorch, which allows configuring the model more precisely.

This project aims to improve the existing NeuralProphet library to allow even more possibilities for its users. Currently, the forecasting model is written on pure PyTorch. It has a complicated structure, code is hardly reusable, making experiments with implementation difficult for outside users. We aim to use the PyTorch Lightning framework to structure the code in a more concise way for future research. Another problem of the current NeuralProphet implementation is that it has a rather specific API regarding initial data format and outputs. NeuralProphet has specific requirements for the model inputs and data preprocessing procedure, which is not the same for other models like ones from PyTorch Forecasting library, for example. Moreover, it has several specific modules, which are not implemented in other models out-of-the-box, as explained in 4.1. All of this makes comparison with other models complicated, as users are required to write additional code to produce comparable results. We aim to introduce state-of-the-art models following the existing API and model output structure. In particular, we aim to create similar to NeuralProphet forecaster classes for each model. This will ensure the same input and output formats, as well as same API to use these additional models. Further, we will rewrite PyTorch Lightning steps of each model to support the NeuralProphet metrics calculation procedure to ensure compatibility.

[*]Equal contribution [1]Skolkovo Institute of Science and Technology, Moscow, Russia. Correspondence to: Polina Pilyugina <polina.pilyugina@skoltech.ru>, Alexey Voskoboinikov <dblokv@gmail.com>.

Final projects of the Foundations to Data Science course 2021 (Skoltech), Skoltech, Moscow, Russian Federation, 2021.

2. Description of the project

2.1. Main goals

Taking into consideration existing drawbacks, we outlined the main goals of our project as follows:

- Refactor the main model class from NeuralProphet with PyTorch Lightning
- Refactor the rest of the code to support PyTorch Lightning in accordance with existing API
- Adapt and include existing implementations of state-of-the-art models for time series forecasting under the NeuralProphet API
- Add hyperparameter tuning with Ray Tune as additional module to NeuralProphet
- Recreate LIBRA framework for benchmarking in Python and run it on NeuralProphet and our additionally included models
- Add neccessary tests and documentation for introduced functional

2.2. Existing solutions

First part of the project is to structure existing code in accordance with existing PyTorch Lightning framework ((Falcon & .al, 2019)). PyTorch Lightning is a lightweight PyTorch wrapper that allows to organise the code into separate PyTorch Lighting modules. This framework provides multiple advantages compared to usual PyTorch: models become hardware agnostic and structured, it provides integration with popular machine learning tools, while keeping flexibility of original PyTorch. Among such tools is Ray Tune for hyperparameter tuning.

PyTorch Lightning also provides a robust architecture that will help us to implement four state-of-the art models for

time series forecasting: N-Beats ((Oreshkin et al., 2019)), LSTM ((Hochreiter & Schmidhuber, 1997)), Temporal Fusion Transformers ((Lim et al., 2019)) and DeepAR ((Salinas et al., 2020)). We rely on existing implementations, available in PyTorch Forecasting library[1].

Another big part of the project is the implementation of additional functionality to the NeuralProphet model. It will include hyperparameter optimization and modules for evaluation. We will use Ray Tune library for hyperparameter optimization. Ray Tune has functional for fast distributed hyperparameter tuning, which allows for additional parallelization and scalability. Moreover, PyTorch Lightning has hooks to Ray Tune, which allow for a seamless connection between the model and Ray Tune. For benchmarking, we will rely on the work on LIBRA framework, described in (Bauer et al., 2021). LIBRA is an evaluation framework that evaluates forecasting methods based on their performance.

All these changes to the NeuralProphet should be accompanied by extensive testing and documentation. NeuralProphet already has modules with integration tests, and we will extend them to cover all of the refactored and new code. As for documentation, we will add necessary documentation for all new functions provided. This will also require adding corresponding pages into the doc of NeuralProphet and adding minimal examples as notebooks to the GitHub repository. We will follow best code style practices and create a modular code with understandable parameter names and docstrings for every function and their parameters.

3. Main Challenges

3.1. API, architecture and style

One of the main challenges in this project is to maintain NeuralProphet style and API. To be more precise, it was inherited from Prophet structure of inputs and outputs. It makes NeuralProphet easily comparable with Prophet, but not with other models. Refactoring in accordance to PyTorch Lightning should not affect existing library consumers. Therefore we will need to introduce additional modules to the TimeNet model used in NeuralProphet, in order to preserve existing output structure and metrics calculation. It possesses some challenges, connected to our goal to maintain all existing functionality in place. We will also refactor existing structure of NeuralProphet to make it modular and understandable.

And we aim to introduce new models in NeuralProphet in accordance with its existing API to preserve usability. This will require us to understand in detail the explicit structure of inputs and outputs to each model and reformat them accordingly. Additionally, we aim to create models in the unified structure, to allow seamless introduction of Ray Tune hyperparameter optimization to all the models.

Existing PyTorch implementations of state-of-the-art models are also available in PyTorch Lightning framework, in PyTorch Forecasting library. However, they do not support NeuralProphet API out of the box, so we will need to refactor them in accordance. In particular, we aim to add metrics calculation and reporting the same way as it is done in NeuralProphet. Also we need to make similar methods and output reporting, in order to allow easy comparison of the results. For this, we will create forecaster classes for each model, that will be similar to the NeuralProphet forecaster class.

3.2. Adapting LIBRA

The other challenge - implement LIBRA framework (Bauer et al., 2021) for our purposes. It's implementation is available[2] only in R, so we will have to adapt it in Python as a part of the NeuralProphet. Further, we will run the benchmarking framework on sample of time series from (Bauer et al., 2021). Main challenges in this case are to obtain the datasets from the R implementation and running the models, which is computationally intense.

4. Implemented functional

4.1. Pytorch Lightning introduction to NeuralProphet

We finished refactoring the NeuralProphet code in the Pytorch Lightning framework. In particular, we implemented all required modules in TimeNet, which is the base model of NeuralProphet. Additionally, we added necessary hooks to connect the model with all the parts of NeuralProphet. Specifically, we preserved the metrics reporting structure and usage of specific optimizers and schedulers provided by the existing NeuralProphet model. NeuralProphet has its specific class for metric calculation called Metrics. It consists of several metrics such as MSE and MAE. This class is supplied with predictions and actual values for each batch on each epoch. On the epoch end, the metrics is computed. In order to introduce this to models, we changed several modules in Pytorch Lightning implementations: training step, validation step, training epoch end, validation epoch end. Additionally, we implemented a module to connect with the hyperparameter optimization functionality we added further. It is for internal use in the hyperparameter optimization module. It initializes the model and all necessary loaders with provided parameters and outputs it without fitting.

The existing library consists of several modules, from which

[1] https://pytorch-forecasting.readthedocs.io/en/latest/models.html

[2] https://github.com/DescartesResearch/ForecastBenchmark

the TimeNet module and forecaster module are the main ones. TimeNet module contains the model itself. We rewrote the model in Pytorch Lightning: we introduced additional required modules, which Pytorch Lightning requires, as well as we introduced necessary hooks for the model to preserve the structure of the forecaster and its modules. Moreover, as was suggested by our colleagues during peer review, we refactored the structure of the code to make it more modular, as will be described in section 7. The forecaster module is the main module used by users, with additional preprocessing and initialization functional. It contains the NeuralProphet forecaste class, which is the main class used by users.

The main methods of NeuralProphet class are:

- init — initializes the class and all the parameters. In case of NeuralProphet, it also initializes metrics structure, which is used further.
- fit — fits the model on the dataset provided. This step consists of preprocessing the data, initializing data loaders, training the model and outputing the training and validation metrics, defined in init module.
- make_future_dataframe — creates a dataframe on which the prediction is done. It contains all the necessary inputs, which the model will require. It support supporting data for prediction on unseen future data, as well as historic prediction.
- predict — predicts the outputs, based on the data, provided in the make_future_dataframe
- plot — plots forecasts in the unified style

On figure 1 is an example of using a NeuralProphet model code structure

```
m = NeuralProphet(
    n_lags=10,
    n_forecasts=3,
    changepoints_range=0.95,
    n_changepoints=30,
    weekly_seasonality=False,
    batch_size=64,
    epochs=10,
    learning_rate=1.0,
)
metrics = m.fit(df, freq='5min')
future = m.make_future_dataframe(df, n_historic_predictions=True)
forecast = m.predict(future)
fig = m.plot(forecast)
```

Figure 1. Example code structure with NeuralProphet model

There are additional methods, but the main changes we made were concerned with these modules. Originally, NeuralProphet included two functions for training_epoch and evaluation_epoch, which we removed and added through Pytorch Lightning into TimeNet. We also moved the main training loop into the Pytorch Lightning Trainer function. Further, it will allow to add new callbacks like EarlyStopping from PyTorch Lightning to NeuralProphet.

4.2. State-of-the-art models

We have added LSTM, NBeats, TemporalFusionTransformer and DeepAR models in accordance with NeuralProphet API.

4.2.1. LSTM

For LSTM, we refactored existing PyTorch implementation in the Pytorch lightning framework. We changed training and validation steps to use the NeuralProphet Metrics class, as in the case of NeuralProphet refactoring. This change ensures that the LSTM model will provide metrics calculated on each epoch the same way NeuralProphet does. We used the existing NeuralProphet forecaster module for the model class and adjusted it to be used with this particular LSTM model. In this case, we made the connection as seamless as possible by changing all NeuralProphet modules to support the new model. Additionally, we added hooks for hyperparameter optimization of LSTM as well. On figure 2 is an example of using a LSTM model code structure

```
m = LSTM(
    n_lags=10,
    n_forecasts=3,
    num_hidden_layers=1,
    d_hidden=64,
    learning_rate=0.1,
    epochs=10,
    batch_size=None,
    loss_func="Huber",
    optimizer="AdamW",
    train_speed=None,
    normalize="auto",
    impute_missing=True,
    lstm_bias=True,
    lstm_bidirectional=False,
)
metrics = m.fit(df, freq="5min")
future = m.make_future_dataframe(df, n_historic_predictions=True)
forecast = m.predict(future)
fig = m.plot(forecast)
```

Figure 2. Example code structure with LSTM model

4.2.2. NBeats

For NBeats we used Pytorch Forecasting NBeats model as a base. We refactored training and validation steps of the model implementation to support the same metrics reporting class as in NeuralProphet. For this we introduced wrappers of predictions, in order for them to be supported by all the metrics inside NeuralProphet metrics class. We implemented additional data preprocessing in order for the

model to have the same input as the NeuralProphet. Models from Pytorch Forecasting Library use specific TimeSeriesDataset class, so we needed to add wrappers to process NeuralProphet inputs into this class. For this we reformulated the TimeSeriesDataset parameters in accordance with NeuralProphet parameters. We used similar initial preprocessing tools for imputing data as are used in NeuralProphet, to induce compatibility. Moreover, the specificity of output processed with PyTorch Forecasting models and its prediction functions required us to completely rewrite main existing modules of NeuralProphet forecaster class, in order to support new model. In particular, by default predictions are stored as dictionaries of specific sort. Therefore in order to calculate metrics for each epoch, we preprocessed the network outputs to be compatible with NeuralProphet Metrics class. Moreover, the prediction function by itself outputs raw results, which we needed to further process and wrap into the same dataset structure, as in case of NeuralProphet. Therefore, we basically introduced a new forecaster holder for the NBeats model, which has the same structure of modules and produces the same outputs as NeuralProphet. It also follows the same logic, as in NeuralProphet: it contains initialization of the model, creation of dataset, fitting and predicting modules. Additionally, Pytorch Forecasting allows to define hyperparameters from the dataset, and we used this functional. On figure 3 is an example of using a NBeats model code structure

```
m = NBeats(
    n_lags=12,
    n_forecasts=3,
    batch_size=None,
    epochs=100,
    num_gpus=0,
    patience_early_stopping=10,
    early_stop=True,
    weight_decay=1e-2,
    learning_rate=3e-2,
    auto_lr_find=False,
    num_workers=3,
)

metrics = m.fit(df, freq = freq)
future = m.make_future_dataframe(df, n_historic_predictions=True)
forecast = m.predict(future)
```

Figure 3. Example code structure with NBeats model

4.2.3. DEEPAR

DeepAR is a probabilistic forecasting model, and Pytorch Forecasting provides a good base implementation. This model uses Normal Distribution Loss (NDL) as a default loss function, and data has to be normalised accordingly (NDL does not work with all-positive target values). As probabilistic model, for NDL it predicts 2 values - ”loc” and ”scale”, and training network with this loss function also decreases other metrics, as shown in the example notebook. We override basic PyTorch Lightning methods (train/validation steps and methods for epoch end), and then wrapped this model to make it work with NeuralProphet datasets and produce metrics and predictions in the same way.

```
m = DeepAR(
        n_lags=32,
        n_forecasts=10,
        batch_size = 32,
        epochs = 10,
        num_gpus = 0,
        patience_early_stopping = 10,
        early_stop = True,
        learning_rate=5e-4,
        auto_lr_find=True,
        num_workers=8,
        hidden_size=10,
        rnn_layers=2,
        dropout=0.1,
)

metrics = m.fit(df, freq = freq)
future = m.make_future_dataframe(df, n_historic_predictions=True)
forecast = m.predict(future)
```

Figure 4. Example code structure with DeepAR model

4.2.4. TEMPORALFUSIONTRANSFORMER (TFT)

Temporal Fusion Transformer (TFT) is a novel architecture, which combines recurrent layers for local processing with self-attention layers for long-term dependencies. We used available in PyTorch Forecasting implementation, wrapping the dataset into TimeSeriesDataset with proper scaling and normalisation of the target. It’s baseline loss function is Quantile Loss, and we introduced it as well. Similar to other models, we override basic functions that are responsible for interacting with PyTorch Lightning, and add functional that mimic public methods of NeuralProphet - fit, make_future_dataframe, predict, and plot. TFT has multiple hyperparameters that determine model - number of hidden layers, number of attention heads etc., so best results can be achieved with hyperparameter tuning.

```
m = TFT(
    n_lags=32,
    n_forecasts=10,
    epochs=10,
    learning_rate=0.03,
    hidden_size=16,
    attention_head_size=1,
    dropout=0.1,
    hidden_continuous_size=8,
)

metrics = m.fit(df, freq = freq)
future = m.make_future_dataframe(df, n_historic_predictions=True)
forecast = m.predict(future)
```

Figure 5. Example code structure with TFT model

In general, all the models we introduced have the same main modules, as NeuralProphet, and output the results in the same format. This allows for a fast comparison of training and validation metrics, as well as plotting of the results with existing NeuralProphet functional. In the future, it is possible to refactor it even further and make implementation easier to read and maintain. We created example notebooks for each of the model, in order to provide users with an introduction to how these models work.

On figure 7 we provide example predictions for three models: NeuralProphet, LSTM, TFT, NBeats and DeepAR, on a sample Yosemite dataset. These baseline results are contained in example notebooks of each model and can be easily rerun, to check different hyperparameter combinations, for example.

4.3. Hyperparameter optimization

As for hyperparameter tuning, we created a separate module in NeuralProphet library for this sake. We added hooks to each of the models we implemented and NeuralProphet itself, using the functionality provided by Pytorch Lightning, to easily connect the optimization module with the main model modules. Moreover, our hyperparameter optimization works not only with NeuralProphet, but for all the models we implemented. Currently, the function works in two modes: auto and manual. The automatic parameter selection uses predefined by us sets of hyperparameters to tune over. The manual mode requires user to provide their own configuration, in accordance with tune API. We have added a notebook example of how this can be done, so the users will have an idea on how to use Ray Tune functionality as well. This function is called tune_hyperparameters and it out-of-the-box requires only three parameters: the model name, the dataframe with data, and frequency of dates in data. Its default set up can be useful for fast basic tuning and also for new users, which have no experience with the model. We evaluated the tuning progress on the Yosemite dataset, included in NeuralProphet example data. We compared the default one-step-ahead forecasting initialization of NeuralProphet with its tuned version, and it showed significant improvement. Therefore, the proposed functional can be exceptionally useful for new users, as it includes automatic mode which requires no initial understanding of the model parameters and provides reasonable improvement out-of-the-box. On figure 6 we provide a resulting metrics of two NP models: with default parameters, and with tuned parameters

We introduced the same functional for all the models we introduced. It works in the same way, as with NeuralProphet. We added an example for LSTM model in the notebook as well.

Ray Tune additionally provides a functionality to distribute resources among trials of experiments which allows for fast training. We used AsyncHyperBandScheduler for hyperparameter tuning. This scheduler is an improved version of HyperBand scheduler. The idea behind it, is that it early stops low-performing trials using the HyperBand optimization algorithm. This allows not to waste computational resources on low-performing hyperparameter configurations, which also induces faster optimization.

SmoothL1Loss	0.005514
MAE	3.886362
MSE	29.934882
RegLoss	0.000000
SmoothL1Loss_val	0.170988
MAE_val	27.786178
MSE_val	928.262373

SmoothL1Loss	0.000469
MAE	1.086094
MSE	3.560099
RegLoss	0.000108
SmoothL1Loss_val	0.001177
MAE_val	1.659725
MSE_val	8.928988

Figure 6. Left is metrics of default model, and right is metrics on tuned model

4.4. Tests and documentation

We have provided each introduced class and function with detailed explanation of parameters. We also have added descriptions on how to use all main functions in README.MD and main docs pages. Moreover, we added example notebooks for all new functional. This notebooks have links to Colab, which allows to test their functional in Colab environment. As for the tests, we have added integration tests for each of the functional we provide, in the same manner, as in original NeuralProphet. By default, they are run automatically when pushing the changes to GitHub repository, if developers version is installed. Or one can run the debug code from tests module, in order to check the current state of the code. More details on how to run these tests, see README.MD in the GitHub repository.

5. LIBRA evaluation

In order to evaluate our results, we used LIBRA framework. We used the LIBRA dataset containing 400 time series from four different usecases:

- Economics — (gas, sales, unemployment, etc.)
- Finance — (stocks, sales prices, exchange rate, etc.)
- Human access — (calls, SMS, Internet, etc.)
- Nature and demographics — (rain, birth, death, etc.)

Following the LIBRA framework, the benchmarking was implemented for different methodologies: one-step-ahead and multi-step-ahead forecasting. One-step-ahead forecasting constitutes to forecasting one period ahead from the set date. Multi-step-ahead forecasting constitutes to forecasting several steps ahead in the future.

We provide results averaged over sample of time series from each domain and each methodology. Overall, the challenge with these datasets was that they contained originally no timestamps, and we needed to review the R implementation to find something resembling actual frequencies. For those datasets, which had some interpretable frequency, we used it. For those without interpretable frequencies, we manually set the frequency to be daily. Moreover, these time series were drastically different in size and while for some of them models ran fast, it was usually not the case. In order to perform the benchmarking we therefore limited number of datasets on which we run the model to a small sample from each domain.

As for the metrics, we have also followed the procedure of original LIBRA framework, and included the following metrics: symmetrical mean absolute percentage error (sMAPE), mean absolute scaled error (MASE), wrong-estimation shares (Mean Under- and Over-Estimation Shares, MUES/MOES) and the mean wrong-accuracy shares (Mean Under- and Over-Accuracy Shares, MUAS/MOAS). Both sMAPE and MASE measures are independent of the scale and can be used across different time series. While MUES/MOES and MUAS/MOAS provide additional insights on whether the model tends to under or over estimate the data. We implemented functions to calculate this errors, so they can also be used in the main NeuralProphet by users. On 1 we provide results of this benchmarking on economics usecase. The best model for multi-step-ahead forecasting is TFT. The best model for one-step-ahead forecasting is NBeats. The results for other usecases can be found in .

In general, performance of NeuralProphet is usually inferior to other models. However, for most other models we used available out-of-the-box in PyTorch Forecasting initialization, which is based on the dataset. And for benchmarking we have only set n_lags, which is fixed in the dataset (as implied frequency); n_forecasts, which is either 1 for one-step-ahead-forecasting or $min(n\_lags, 0.2 * len\_ts)$. The other parameters were learning_rate and number of epochs, which were also the same for all models on all time series. The main difference is that for NBeats, DeepAR and TFT, the implementation uses the default function to infer parameters based on the time series, therefore they were better adjusted than NeuralProphet. Although NeuralProphet has some parameter adjusting, it is still a work in progress, which will likely be improved further.

As for the code, we implemented a module to run the LIBRA benchmarking. Example of its usage is available in LIBRA notebook in example notebooks section of our repository. Note, that even small number of time series in benchmarking is still computationally intense, as it includes deep models. We also provide a notebook to combining results into nice LaTex tables, as can be seen in evaluation results notebook in our repository.

metrics	method	DeepAR	LSTM	NBeats	NP	TFT
mase	MSA	**1.95**	26.89	2.73	66.78	55.42
moas	MSA	0.07	0.06	0.05	0.64	**0.02**
moes	MSA	0.39	0.19	0.32	0.41	**0.09**
muas	MSA	**0.05**	0.30	0.06	1.29	0.71
mues	MSA	0.61	0.81	0.68	**0.59**	0.91
smape	MSA	10.54	46.34	**10.44**	2281.54	131.75
mase	OSA	2.97	29.45	**2.23**	61.12	11.72
moas	OSA	0.06	**0.03**	0.06	0.76	0.07
moes	OSA	0.35	**0.14**	0.33	0.53	0.29
muas	OSA	**0.06**	0.37	**0.06**	0.54	0.46
mues	OSA	0.65	0.86	0.67	**0.47**	0.71
smape	OSA	11.33	57.07	**10.71**	316.01	91.61

Table 1. Results of benchmarking on economics usecase

6. A link to the GitHub repository

For this project we forked the original repository and will contribute to it. It will allows us to create a pull request into the master in the future. GitHub repository is available through https://github.com/adasegroup/neural_prophet/tree/master. We have preserved the initial GitHub repository structure. On figure 8 we provide an overview of GitHub repository structure. We outlined the files that were added or changed on the course of this project. This includes: refactoring of existing main modules and structure, documentation, notebooks and tests.

7. Peer review suggestions

We carefully read all peer reviews, and particularly stopped on the review by Cohortney team. We are very grateful to the colleagues, as their suggestions on improving the code structure were very helpful. The team suggested to make the structure modular. Original NeuralProphet, and our re implementation were of the same structure, initially. However, the original code structure of NeuralProphet indeed lacked modularity and it was extremely hard to read the code, not to say about refactoring it and understanding. We also believe, that the original structure was too chaotic, so we took the advice of colleagues and their review and refactored the code structure as well. Currently, we have made separate modules in NeuralProphet with forecasters, models, tools, metrics etc. We also refactored initialization files so that each new model is imported directly from NeuralProphet, and not from some of its modules. There was also a suggestion by the team to move trainer calling on the model from forecaster, however, we wanted to maintain the original API and thus having train in forecaster class was necessary. Overall, this was a great input which, in our opinion, allowed us to harshly improve understandability of

NeuralProphet structure.

8. Prospects for further work

During the project, we deeply understood the structure and features of NeuralProphet. And although we did a hard job on its improvement, there are still areas in which future work could be conducted. First and the most important, in our opinion, is further refactoring to induce its performance on GPU, as currently not all the parts of NeuralProphet are GPU-scalable. Moreover, PyTorch Lightning allows to add new functional callbacks to NeuralProphet. Secondly, we have seen on the evaluation, that other models are behaving better without hyperparameter tuning. We have added the hyperparameter optimizator as a possibility for NeuralProphet, but in future it will be useful to add more sophisticated initial parameter set up, alike in PyTorch Forecasting models. Another useful insight we encountered, when working with non-time-stamped data from Libra. NeuralProphet is currently not timestamp-agnostic, however it might be exceptionally useful in some cases, where underlying timestamp is not defined or does not make sense. Therefore, we look forward to continue working on it in the future, as we see a potential behind this library.

References

Bauer, A., Züfle, M., Eismann, S., Grohmann, J., Herbst, N., and Kounev, S. Libra : A Benchmark for Time Series Forecasting Methods Libra : A Benchmark for Time Series Forecasting Methods. (April), 2021.

Falcon, W. and .al. Pytorch lightning. *GitHub. Note: https://github.com/PyTorchLightning/pytorch-lightning*, 3, 2019.

Hochreiter, S. and Schmidhuber, J. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997. ISSN 08997667. doi: 10.1162/neco.1997.9.8.1735.

Lim, B., Arık, S., Loeff, N., and Pfister, T. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *arXiv*, (Bryan Lim):1–27, 2019.

Oreshkin, B. N., Carpov, D., Chapados, N., and Bengio, Y. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. *arXiv*, pp. 1–31, 2019. ISSN 23318422.

Salinas, D., Flunkert, V., Gasthaus, J., and Januschowski, T. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3):1181–1191, 2020. ISSN 01692070. doi: 10.1016/j.ijforecast.2019.07.001. URL https://doi.org/10.1016/j.ijforecast.2019.07.001.

Taylor, S. J. and Letham, B. Forecasting at Scale. *American Statistician*, 72(1):37–45, 2018. ISSN 15372731. doi: 10.1080/00031305.2017.1380080.

A. Figures

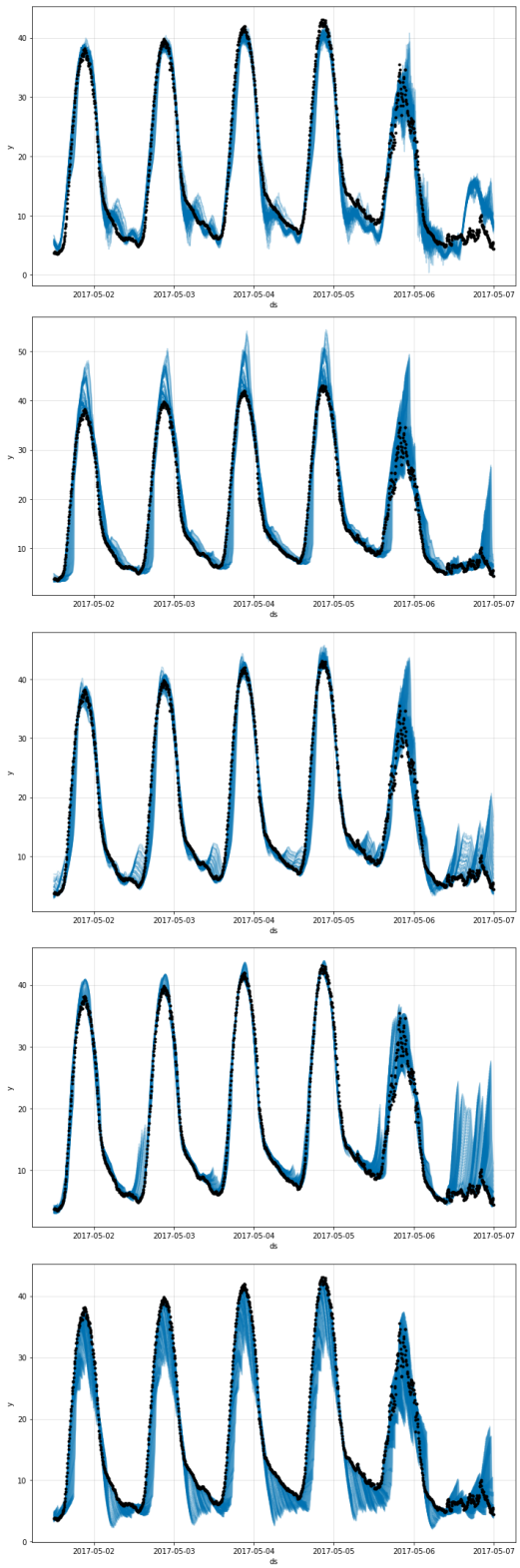

Figure 7. Example of predictions plotting for models (from top to the bottom): NeuralProphet, NBeats, LSTM, TFT and DeepAR

```
└── neural_prophet
    ├── README.md
    ├── docs
    │   ├── images
    │   ├── model
    │   │   ├── additional_models.md
    │   │   └── hyperparameter_optimization.md
    │   └── zh
    ├── example_data
    │   └── LIBRA
    ├── example_notebooks
    │   ├── DeepAR_example.ipynb
    │   ├── LIBRA.ipynb
    │   ├── LSTM_example.ipynb
    │   ├── NBeats_example.ipynb
    │   ├── TFT_example.ipynb
    │   ├── evaluation_results.ipynb
    │   └── hyperparameter_optimization_example.ipynb
    ├── neuralprophet
    │   ├── dataset
    │   ├── forecasters
    │   │   ├── forecaster.py
    │   │   ├── forecaster_DeepAR.py
    │   │   ├── forecaster_LSTM.py
    │   │   ├── forecaster_NBeats.py
    │   │   ├── forecaster_NBeats_old.py
    │   │   └── forecaster_TFT.py
    │   ├── hyperparameter_tuner.py
    │   ├── libra.py
    │   ├── models
    │   │   ├── DeepAR.py
    │   │   ├── LSTM.py
    │   │   ├── NBeats.py
    │   │   ├── TFT.py
    │   │   └── time_net.py
    │   ├── tools
    │   │   ├── configure.py
    │   │   ├── hdays.py
    │   │   ├── metrics.py
    │   │   ├── metrics_libra.py
    │   │   ├── plot_forecast.py
    │   │   └── plot_model_parameters.py
    │   └── utils
    │       ├── df_utils.py
    │       ├── utils.py
    │       └── utils_torch.py
    ├── notes
    ├── peer_reviews
    ├── reports
    ├── requirements.txt
    ├── results_benchmarking
    │   └── example
    │       ├── results_libra_multistep_economics.csv
    │       ├── results_libra_multistep_finance.csv
    │       ├── results_libra_multistep_human.csv
    │       ├── results_libra_multistep_nature.csv
    │       ├── results_libra_onestep_economics.csv
    │       ├── results_libra_onestep_finance.csv
    │       ├── results_libra_onestep_human.csv
    │       └── results_libra_onestep_nature.csv
    ├── roadmap_gantt.png
    ├── scripts
    ├── setup.py
    └── tests
        ├── debug.py
        └── test_integration.py
```

Figure 8. Structure of GitHub repository with outlining the files added or changed during the project, as for the current state

B. Tables

metrics	method	DeepAR	LSTM	NBeats	NP	TFT
mase	MSA	3.22	7.05	2.85	30.9	11.44
moas	MSA	0.38	0.27	0.35	0.75	0.07
moes	MSA	0.53	0.34	0.52	0.38	0.17
muas	MSA	0.12	0.27	0.09	1.72	0.56
mues	MSA	0.47	0.66	0.48	0.62	0.83
smape	MSA	26.43	44.7	24.92	591.34	105.08
mase	OSA	-	-	-	-	-
moas	OSA	0.43	0.33	0.38	0.82	0.32
moes	OSA	0.53	0.36	0.53	0.52	0.13
muas	OSA	0.06	0.18	0.07	0.39	0.45
mues	OSA	0.47	0.64	0.47	0.48	0.87
smape	OSA	24.66	37.48	24.04	539.62	94.61

Table 2. Results of benchmarking on finance usecase. MSA method for multi-step-ahead, OSA for one-step-ahead

metrics	method	DeepAR	LSTM	NBeats	NP	TFT
mase	MSA	1.05	1.67	0.5	5.21	2.81
moas	MSA	0.27	0.24	0.07	0.84	0.08
moes	MSA	0.46	0.33	0.56	0.54	0.34
muas	MSA	0.1	0.17	0.05	0.61	0.39
mues	MSA	0.54	0.67	0.44	0.46	0.66
smape	MSA	15.11	22.16	6.12	113.25	74.52
mase	OSA	-	-	-	-	-
moas	OSA	0.04	0.08	0.03	0.07	0.02
moes	OSA	0.43	0.39	0.48	0.21	0.27
muas	OSA	0.04	0.08	0.03	0.58	0.37
mues	OSA	0.57	0.61	0.52	0.79	0.73
smape	OSA	7.93	14.94	5.74	180.59	72.83

Table 3. Results of benchmarking on nature usecase

metrics	method	DeepAR	LSTM	NBeats	NP	TFT
mase	MSA	-	-	-	-	-
moas	MSA	0.73	0.73	0.28	4.26	0.2
moes	MSA	0.57	0.57	0.44	0.57	0.49
muas	MSA	0.16	0.23	0.24	1.9	0.22
mues	MSA	0.43	0.43	0.56	0.43	0.51
smape	MSA	58.49	59.61	51.59	1206.33	34.42
mase	OSA	-	-	-	-	-
moas	OSA	0.13	0.32	0.1	0.53	0.1
moes	OSA	0.55	0.56	0.44	0.43	0.41
muas	OSA	0.06	0.19	0.03	1.06	0.2
mues	OSA	0.45	0.44	0.56	0.57	0.59
smape	OSA	33.72	34.95	17.82	492.24	52.7

Table 4. Results of benchmarking on human usecase

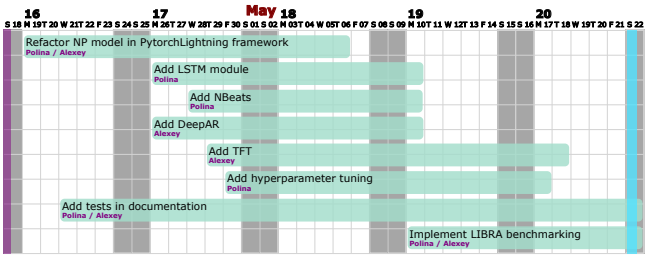

Figure 9. Roadmap of the project with Role Distribution

C. The roles for the participants

We distributed our main tasks and goals evenly, as described on the 9. Both of us were working on refactoring into PyTorch Lightning. Alexey focused on the main TimeNet model class, while Polina worked on the forecaster code. We also distributed addition of models, such that Polina implemented on N-Beats and LSTM, while Alexey implemented Temporal Fusion Transformers and DeepAR. We both have writen corresponding tests and documentation of implemented modules. Further, Polina focused on hyperparameter tuning addition, while Alexey implemented LIBRA framework in python. Afterwards, we both worked on the benchmarking using LIBRA framework and finalization of the project.

At a current stage, we have finished all the aims that were set for the project. The only bottleneck that we encountered was the computational intensity of benchmarking, therefore we completed benchmarking only on sample from all time series available.