Univariate and multivariate time-series methods to forecast dairy income

Vahid Naghashi, Gabriel M. Dallago, Abdoulaye Banire Diallo, Mounir Boukadoum

Departement d'Informatique, UQAM, Montreal, QC, Canada
PO Box 8888 Downtown station, Montreal, QC H3C 3P8, Montreal, Canada
diallo.abdoulaye@uqam.ca

Abstract

Forecasting the income from milk sales can be addressed as a time-series problem since the sequence of multiple dairy attributes during lactation cycles are inter-related and temporally dependent. In this paper, we provide a framework to forecast the income from milk sales during the third lactation of the dairy cows based on dairy attributes recorded through the first and second lactation. We modeled the problem as univariate and multivariate time-series predictions. We propose several state-of-the-art implementations with ARIMA, N-BEATS, transformer and an original method, MuMu+attention, that combines Long-Short Term Memory neural network and attention mechanism to capture the temporal dependencies. To benchmark the implemented methods, we curated data from 147,749 dairy cows from 5,844 Canadian herds. The monthly income from milk sales ($CAD) measured at each cow during their third lactation was treated as the prediction target. The dataset was composed of dairy attributes of milk quality, production, season, year, and health recorded over the first and second lactation of the dairy cows. The results highlighted that most of the methods can achieve relative good performance with the best prediction accuracy obtained by MuMu+attention. MuMu+attention results were 43% better over the classic ARIMA model. By forecasting the income from milk sales, our model could help farmers to early identify less profitable animals and better allocate resources.

Introduction

Dairy farming is one of the largest sectors in agriculture that has been under study through data-driven and machine learning methodologies. Promising results were obtained in automatic cropping of the cow's body region and cow's pattern identification for individual animals (Zin et al. 2018). In another work, deep convolutional neural networks were used for detection of the key parts of a dairy cow's body, resulting in an accurate detector (Jiang et al. 2019). Promising results were obtained by exploiting a deep learning model for calving prediction from activity, lying, and ruminating behaviors of dairy cattle (Borchers et al. 2017). Therefore, the use of machine learning and deep learning based models lead to promising results and its application to predict dairy production could yield satisfying results (Frasco et al. 2020).

Income from milk sale is the main factor associated with the profitability of a dairy farm. A forecasting tool would allow farmers to optimize the allocation of resources by early identifying and removing less profitable animals from the herd. The problem of forecasting income from milk sales can be modeled as a univariate time-series task since the lactation cycles are inter-related and temporally dependent, or multivariate since milk production, the consequently income, is affected by heath, productivity, environmental, and management conditions. Among the classical time-series prediction methods, Autoregressive Integrated Moving Average (ARIMA) has shown a good performance in univariate time-series prediction tasks (Contreras et al. 2003). Recently, deep learning models have been exploited in time-series prediction domain. N-BEATS (Oreshkin et al. 2019) is one of the well-known deep prediction models, in which residual connections are used for univariate time-series prediction and the model's architecture is based on a very deep stack of fully-connected layers. In MuMu (Frasco et al. 2020), Long-Short Term Memory (LSTM) network was exploited in the dairy forecasting field and gave rise to auspicious results.Recently, Transformer models (Vaswani et al. 2017) got more attention of researchers due to their ability to represent the long-term temporal dependencies efficiently by incorporating multi-head self attention in their structure. The efficacy of the self-attention layers gave us the hint about using an attention module in our prediction framework.

The objective of this paper is twofold: first, we propose an extension of MUMU using LSTM and an attention mechanism to forecast the income from milk sales; second, we benchmark univariate models against multivariate ones in predicting future profit using a well curated data from 147,749 dairy cows.

Preliminary

The problem of lifetime milk revenue prediction can be posed as follows. For each input example (ith cow sample) of length T, i.e. $x_i = (x_i^1, ..., x_i^T) \epsilon \Re^{p \times T}$ with p as the number of input dairy factors and T the length of the time-series (total length of the first and second lactation fixed for all the cow samples), a prediction model forecasts the upcoming milk revenues of M steps (months) ahead in third lactation, $\hat{r_i} \epsilon \Re^M$. Therefore, the goal is to learn a function f which maps the input multivariate time-series $X \epsilon \Re^{s \times T \times p}$ to the estimated milk income values in the future lactation months

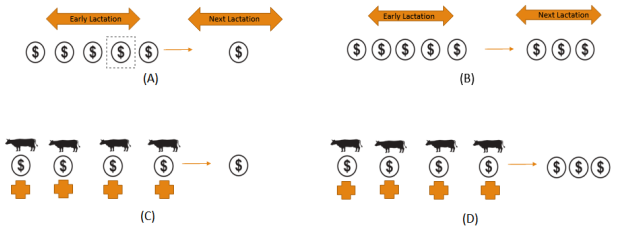

Figure 1: Univariate and multivariate statement of our dairy prediction problem.

$\hat{r} \epsilon \Re^{s \times M}$, with s being the number of the cow samples presented to the model as the training set: $\hat{r} = f(X)$.

Methodology

Prediction of dairy income can be stated as four different cases (Figure 1): univariate or multivariate inputs, single or multiple outputs. In the first case, the input is a sequence of input dairy incomes during the early lactation cycles and the output is the value of the next income in the next lactation. The second case is similar to the first one, with the difference that the output is multiple dairy incomes. In the third case, the input window consists of multivariate series associated with the multiple dairy factors through the early lactation and the output window corresponds to a single income in the next lactation. Finally, the input window in the fourth case contains multivariate series of multiple dairy attributes similar to the latter case, and have a sequence of dairy incomes as its outputs.

Model Architecture

Here we propose MuMu+Attention, which builds on top of the MuMu model (Frasco et al. 2020). It implements time-series of the individual dairy attributes corresponding to earlier lactations. It consists of two LSTM layers stacked on top of each other and one attention layer, followed by a linear layer acting as a decoder. The architecture of the proposed framework is illustrated in Figure A1.

For the purpose of using all the hidden states and temporal feature selection, an attention layer (Vaswani et al. 2017) has been embedded on top of the last LSTM layer in our model. The attention layer consisted of two linear layers and one $tanh$ activation function in between. The final attention weights were the result of applying a $softmax$ function to the output of the second linear layer to normalize the attention scores. These weights were multiplied by the corresponding hidden states and a weighted hidden vector was calculated in order to be fed into the final linear layer which generates the final prediction over the target window.

The corresponding formulas describe the components of the attention layer:

$$L_1 = tanh(W_1 x + b_1) \quad (1)$$

$$Attention\_Outputs = Softmax(W_2 L_1 + b_2) \quad (2)$$

In the above formulas, x represents the output of the last LSTM layer with the shape of (batch_size, sequence_length, hidden_size). $W1$, $W2$, $b1$ and $b2$ are learnable parameters which have been trained in an end to end manner together with the other parameters of the model.

The major advantage of using the attention layer is that the information in all of the hidden states has been exploited, rather than using only the output of the last LSTM cell and this can be referred to as a temporal feature selection step which learns and assigns the importance weights to the sequence of the hidden states associated with the input window (input time steps). A dropout layer is used in the output of the second LSTM layer in order to avoid over-fitting. The value of the dropout rate hyper-parameter determines the ratio of the hidden states whose outputs are dropped out and this hyper-parameter is set to 0.5 in our model. Finally, the decoder in our model, which is a linear layer with no activation function, generates the output predictions by one forward pass instead of the time consuming dynamic decoding used in the conventional architectures. In our work, the mean squared error was used as the loss function while training the model.

Experimental settings

Data

The input to our prediction model was a multivariate time-series containing a set of dairy attributes, including metrics of milk quality, seasonality, year, health, and management factors, recorded during the first and second lactation of 147,749 dairy cows from 5,844 Canadian dairy farms over the years of 2006 to 2017. The prediction targets were the monthly income from milk sales ($CAD) measured at each cow during their third lactation.

Data Pre-processing

In the experiments, lactation length was fixed to 11 months for first, second, and third lactation based on the mean + one standard deviation of the lactation lengths related to the training cow samples. For cows with lactation lengths shorter than 11 months, additional data rows were created in the missing months and imputed through linear interpolation based on the two closest months. The following steps were taken to clean the data: 1- Keeping only animals having test records in the first, second and third lactations, 2- Deleting the records from the dry period: dry period is defined as the months in which a cow doesn't produce milk (which mostly occurs between two lactation cycles) and milk value (income) is almost zero. Since there is no income from sales during the dry period, records associated with these months were deleted, 3- Removing the animals who left their herd before and during the third lactation, 4- Deleting duplicate records, 5- Deleting the records with negative milk value and cumulative milk value: The reason for removing such records, which constituted a small percentage of the data, was to remove the inconsistencies in the data set, as the milk value is the income from milk sale (CAD) and negative values of dairy income could be accounted for as an inconsistency due to the errors in data acquisition steps., 6- Deleting the records which included contradictions, e.g., rows indicating the cow was milking, while the milk yield of those

records was equal to zero, 7- Outlier removal (their corresponding rows): The values outside the range of Mean -/+ 2.5 $\times$ standard deviation were specified as outliers for a given dairy attribute and were deleted. All the dairy attributes used as input to our model and their descriptions are shown in Tables A1 and A2 (Appendix).

Among the 147,749 dairy cows, 100,000 were selected to train and the remaining 47,749 cow samples to test the models. Missing data was imputed after train-test split to avoid information leakage (Thomas et al. 2020). In our work, missing value imputation were performed in 7 consecutive steps: first the cow samples were grouped based on their herd ID, season and year, then the missing values within each group were imputed using the average of non-missing samples within the same group. The imputation process in the remaining six steps was similar as in the first one, with the following attributes used for grouping, respectively: (herd ID, season), (herd id, year), (herd id), (season, year), (season), (herd id).

Experimental Details

Baselines We choose five forecasting methods as comparison models: two univariate (ARIMA (Contreras et al. 2003) and N-BEATS (Oreshkin et al. 2019)) and two multivariate models (MuMu (Frasco et al. 2020) and a standard Transformer model (Vaswani et al. 2017) adapted for time-series prediction. Most of the above methods are well known in the time-series domain. The ARIMA model is chosen as a baseline method as it is a classic statistical approach in time-series analysis. The parameters used in the ARIMA are taken from (Frasco et al. 2020) which used a stepwise algorithm to determine p, q, P and Q parameters. The $N-BEATS$ model (Oreshkin et al. 2019) had four blocks. The first two blocks had a trend basis function in their outputs, and the other two contained a seasonality basis. The number of fully-connected layers inside each block was fixed to 4 with hidden-size of 128. As the N-BEATS model is designed to receive a univariate sequence, we fed the sequence of milk incomes in 22 months as its input. The parameters for the MuMu model are: 2 LSTM layers, one linear layer (without an activation function) and the hidden size is fixed to 32 in all layers. The Transformer model which was used as another baseline method in our experiments, was composed of two encoders and one decoder layer, in which the input of the decoder was the last time step of the input window.

Grid Search for Hyper-parameters determination Grid search was conducted to determine the hyper parameters (batch size, learning rate and hidden size of the LSTMs) due to its common usage in other related works and satisfying results (Zhou et al. 2021). Our proposed model was optimized with Adam optimizer and learning rate of $1e^{-4}$. The total number of epochs was set to 20 with an early stopping based on the validation loss. **Setup:** All the numerical inputs were standardized with zero mean and unit standard deviation. At the same time, categorical variables were normalized to the range between 0 and 1. The above transformation was applied to each dairy feature, separately.

RMSE results				
Univariate		Multivariate		
ARIMA	N-BEATS	Transformer	MuMu	MuMu+Attention
7.094	4.198	4.064	4.082	**4.052**

Table 1: Prediction results in terms of RMSE (teste dataset).

MAE results				
Univariate		Multivariate		
ARIMA	N-BEATS	Transformer	MuMu	MuMu+Attention
5.560	3.256	3.178	3.175	**3.143**

Table 2: Prediction results in terms of MAE (test dataset).

The input window was set to 22 (with the length of 11 for both of the first and second lactation). The prediction (target) window size (length of the third lactation) was fixed as 11 in our experiments according to the mean of the lactation lengths corresponding to all the cow samples. We evaluated our prediction framework using the $RMSE = \sqrt{\frac{1}{N_{cows} \times N_{months}} \sum_{m \epsilon months} \sum_{c \epsilon cows} (\hat{p}_{c,m} - p_{c,m})^2}$ and the $MAE = \frac{1}{N_{cows} \times N_{months}} \sum_{m \epsilon month} \sum_{c \epsilon cows} |\hat{p}_{c,m} - p_{c,m}|$ on the forecasting window. All the models were trained and tested on 4 NVIDIA T4 Turing GPUs with 16 GB GDDR6 memory. Additionally, multiple pairwise Wilcoxon tests, with Bonferroni p-value adjustment, were used to compare models.

Results and Discussion

Tables 1 and 2, summarize the forecasting accuracy of all the methods on the test dataset. The best results are highlighted in boldface in each setting. We also reported the chosen hyper-parameters (the best combination), including hidden size, learning rate and batch size after performing the grid search in Table A3 (Appendix).

MuMu+Attention model was able to forecast the income with the highest accuracy based on the RMSE (Table 1) and MAE (Table 2) measures. The LSTM layers included were able to represent and capture the long-term temporal dependencies (Gers, Schmidhuber, and Cummins 2000) by exploiting the input, forget, update, and output gates in its structure. Those gates help the LSTM to select the most relevant information and update the previous state using the current input at each time step. At the same time, forget and update elements give the LSTM the capability of remembering the long-range dependencies and making use of such relationships in the prediction process.

Among the baseline methods, the classical ARIMA forecasts had the highest error. The linear nature of this model hinders its ability to capture more complex and non-linear temporal correlations. The N-BEATS model had a relatively higher error compared to the multivariate approaches, which is likely because it does not have a sequential module in its structure to represent the temporal dependencies. Capturing long-range dependencies is a key factor in time-series prediction (Zhou et al. 2021) and all the LSTM or Transformer based approaches try to represent such relations by captur-

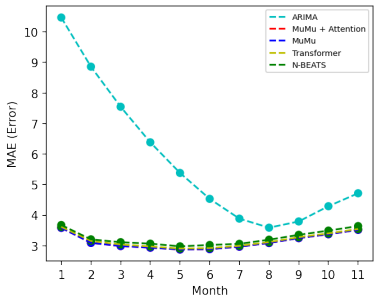

Figure 2: Monthly Errors (MAE) of different methods and our proposed framework on prediction of the milk incomes.

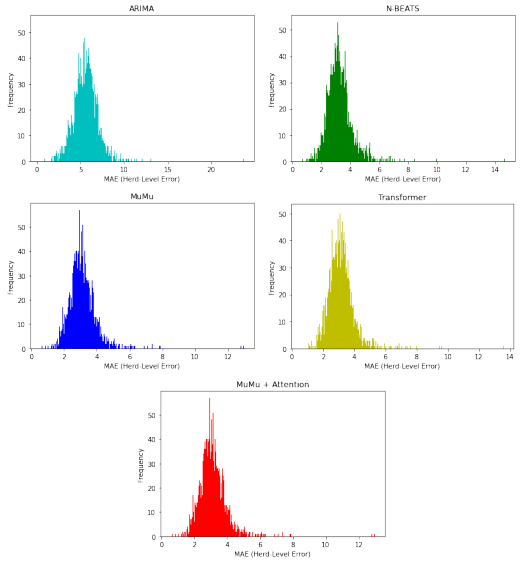

Figure 3: Herd-based Error (MAE) distribution of different methods.

ing the impact of different time steps on each other, which seems to be a dominant factor for the success of those methods. The reason behind the relatively lower performance of the Transformer model might be due to the lack of sufficient and enough data for training of this model, which has more parameters than the other models (Table A7). As the errors related to different deep learning models are in the same range, the training and inference time of the MuMu+Atten and other models are reported in table A4 to give an insight on the time complexity of the deep learning based methods. Based on the results, the N-BEATS model needs more training and inference time than the MuMu+Atten, despite its lower performance compared to the multivariate models. We further conducted other experiments to predict the milk value in a single month after the first and second lactation (input window). The results indicate that the model nearly gave the same results as the prediction of the multiple months (Table A5).

Figure 2 depicts the forecast accuracy of the proposed framework in comparison to the other models over each month of the third lactation. Except for the ARIMA model, the best forecasts occurred in the middle of lactation (month 5) compared to the beginning and the end. This is likely because there is a great variability among cows in the amount of milk produced and consequently sold during these steps, making it more difficult to forecast. The distribution of MAE for all models at the herd level is plotted in Figure 3. There was no strong evidence of herd-bias as the distributions were visually similar. The MAE, at the herd level, was not statistically different between the multivariate models ($p \geq 0.05$) and it was lower than both univariate models ($p < 0.05$; Table A6). This indicated that the proposed framework could also capture the long-range temporal dependencies in input and output windows by processing the input sequence using the LSTM layers and representing the importance of each time step through the attention weights. Furthermore, the generative style decoder (the output linear layer in our framework) acquired the output predictions in one forward pass and avoided the error accumulation during the testing phase.

Conclusion

In conclusion, we studied the problem of forecasting the income from milk sales by designing a framework and comparing univariate and multivariate approaches. We enriched our framework with the MuMu+Attention model that combines LSTM and attention mechanism. The experimental results showed that multivariate models tend to perform better even though the performance of NBEAST can be an important way of approximating the profit just using a univariate time-series. However, we showed that MuMu+Attention provided the highest accuracy.

References

Borchers, M.; Chang, Y.; Proudfoot, K.; Wadsworth, B.; Stone, A.; and Bewley, J. 2017. Machine-learning-based calving prediction from activity, lying, and ruminating behaviors in dairy cattle. *Journal of dairy science*, 100(7): 5664–5674.

Contreras, J.; Espinola, R.; Nogales, F. J.; and Conejo, A. J. 2003. ARIMA models to predict next-day electricity prices. *IEEE transactions on power systems*, 18(3): 1014–1020.

Frasco, C. G.; Radmacher, M.; Lacroix, R.; Cue, R.; Valtchev, P.; Robert, C.; Boukadoum, M.; Sirard, M.-A.; and Diallo, A. B. 2020. Towards an Effective Decision-making System based on Cow Profitability using Deep Learning. In *ICAART (2)*, 949–958.

Gers, F. A.; Schmidhuber, J.; and Cummins, F. 2000. Learning to forget: Continual prediction with LSTM. *Neural computation*, 12(10): 2451–2471.

Jiang, B.; Wu, Q.; Yin, X.; Wu, D.; Song, H.; and He, D. 2019. FLYOLOv3 deep learning for key parts of dairy cow body detection. *Computers and Electronics in Agriculture*, 166: 104982.

Oreshkin, B. N.; Carpov, D.; Chapados, N.; and Bengio, Y. 2019. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. *arXiv preprint arXiv:1905.10437*.

Thomas, R. M.; Bruin, W.; Zhutovsky, P.; and van Wingen, G. 2020. Dealing with missing data, small sample sizes, and heterogeneity in machine learning studies of brain disorders. In *Machine learning*, 249–266. Elsevier.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; and Zhang, W. 2021. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 11106–11115.

Zin, T. T.; Phyo, C. N.; Tin, P.; Hama, H.; and Kobayashi, I. 2018. Image technology based cow identification system using deep learning. In *Proceedings of the International MultiConference of Engineers and Computer Scientists*, volume 1, 236–247.

Appendix

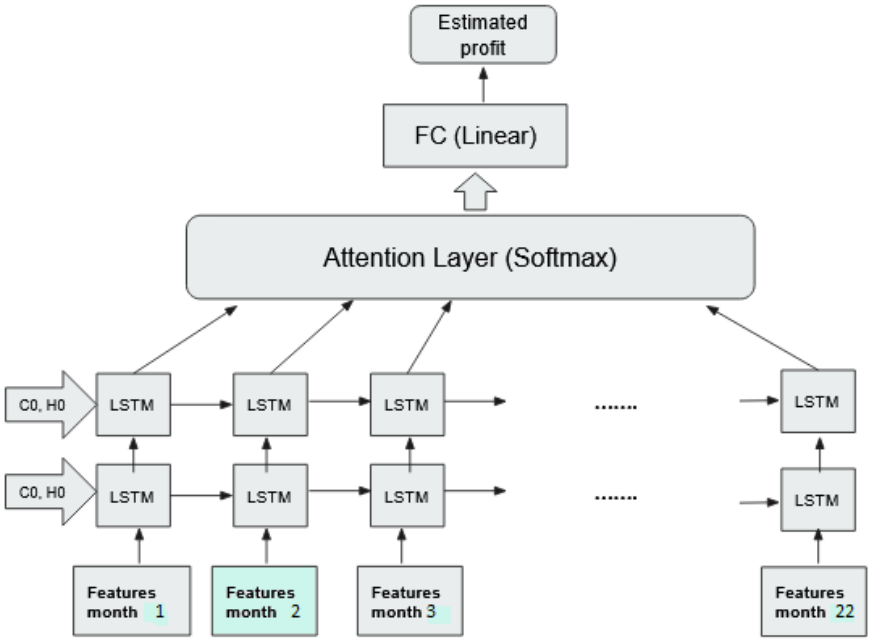

Figure A1: Architecture of the proposed prediction framework.

Variable	Missing %	Min	Max	Mean	Std
Milk produced in 24 hours (Kg)	1.4	0.20	61.40	29.60	9.25
Milk income (CAD)	0.002	0.09	42.95	21.21	5.94
Somatic Cell Count (1000/mL)	17.62	1	1754	156.59	254.87
Milk Urea Nitrogen	56.04	2.10	19.80	10.91	3.27
Lactose yield (Kg)	37.63	0	6.30	4.55	0.22
Fat yield in 24 hours (Kg)	1.64	0	2.40	1.19	0.38
Protein yield in 24 hours (Kg)	1.64	0	1.94	1.00	0.30
Number of days cow has been in milking	0	0	517	169.56	99.39
Fat to protein ratio	1.64	0.73	1.67	1.20	0.16

Table A1: Numerical dairy variables used in our framework.

Categorical variables				
Variable	**Missing %**	**Levels**	**Number of observations**	**% of each level**
Lactation Number	0	1	1625239	33.33
		2	1625239	33.33
		3	1625239	33.33
Number of milking per day	0	1	163201	3.33
		2	4650067	95.37
		3	62449	1.28
Test season	0	1	1235340	25.33
		2	1226102	25.14
		3	1220232	25.03
		4	1194043	24.49
Birth season	0	1	1109460	22.75
		2	1325577	27.18
		3	1274922	26.14
		4	1165758	23.91
Animal condition	0	2	4846703	99.40
		4	29014	0.06
Test year	0	6	155074	3.18
		7	304329	6.24
		8	443724	9.10
		9	486607	9.98
		10	472280	9.68
		11	479705	9.83
		12	498897	10.23
		13	511067	10.48
		14	521706	10.70
		15	486234	9.97
		16	331883	6.81
		17	170715	3.50

Table A2: Categorical dairy variables used in our framework.

Hyper-parameter	Selected value
Hidden size	32
Learning rate	$1e^{-4}$
Batch size	1
Epochs	20

Table A3: The selected hyper-parameters in training of our proposed model

ARIMA	N-BEATS	Transformer	MuMu	MuMu+Atten
$>$ 12 hours	4 hours, 7 mins	2 hours, 34 mins	1 hour, 36 mins	2 hours
$O(L * p^3)$	$O(c * L * d)$	$O(L^2 * d)$	$O(L * d^2)$	$O(L * d^2)$

Table A4: Running time of different models (Training + Inference time). In this table, L is the length of the time series, d in the model's dimensionality, c is a multiplier, and p is the order of the ARIMA model.

	One month RMSE	One month MAE
N-BEATS	4.72	3.72
Transformer	4.58	3.60
MuMu	4.57	3.62
MuMu+Attention	4.56	3.61
ARIMA	13.53	11.81

Table A5: RMSE and MAE of the forecasting the milk income (value) in a single month ahead of the input window (first and second lactation)

	ARIMA	N-BEATS	Transformer	MuMu
N-BEATS	< 0.001			
Transformer	< 0.001	< 0.001		
MuMu	< 0.001	< 0.001	0.05	
MuMu+Attention	< 0.001	< 0.001	0.05	1.00

Table A6: Pairwise comparisons of the mean absolute errors for all models at the herd level using the Wilcoxon rank sum test. The Bonferroni method was used to adjust the p-values for multiple comparisons.

Model Name	Number of parameters
N-BEATS	109325
Transformer	276587
MuMu	15339
MuMu+Attention	16428
ARIMA	3

Table A7: Number of parameters of different models