PREDICTION OF POSTPAID CUSTOMERS ELECTRICITY SALES USING LSTM AND SARIMA

Paulus Saritosa - 23522315 (*Author*)
Program Studi Magister Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Jalan Ganesha 10 Bandung
E-mail (gmail): 23522315@std.stei.itb.ac.id

***Abstract*— Sales are an important element in a company. PLN as an electricity company of course also depends on the sale of electricity. One of the products sold by PLN is postpaid products. This research evaluates two forecasting methods, namely Long Short Term Memory (LSTM) and Seasonal Auto-Regressive Moving Average (SARIMA). The data used is electricity sales data for postpaid customers from January 2017 to December 2023. The LSTM and SARIMA forecasting process is carried out by forming the best model using several parameters. From the test results it was found that the SARIMA model was better than LSTM where the RMSE figure was 107.85 and MAPE was 5.7. This research provides valuable insight into selecting the right model for predicting postpaid customers, because LSTM is more effective when the data is large.**

Keywords—LSTM, SARIMA, Electricity Sales

I. Introduction

The State Electricity Company (PLN) is a company operating in the electricity services sector, which is oriented towards customer service while still paying attention to company profits. The company's income is obtained from various aspects, and the main contributor is from selling electricity to customers. Electricity sales have a big influence on the company's income, because from this income the company can manage finances that can be budgeted for the company's needs or requirements. PLN's main products are prepaid and postpaid products, where prepaid means purchasing tokens first and then enjoying electricity, while postpaid is the opposite. The current problem is that there is no appropriate method used by PLN to carry out forecasting. Forecasting electricity sales has a direct impact on company revenues and sales, and it can be an important basis for determining the next period's budget and setting performance targets for the next period. Forecasting electricity sales can provide a clearer view of the expected revenues and costs that a company will face. This enables better planning, more informed decision making in managing Company resources and operations and can improve electricity services to customers.

Forecasting can be classified based on time, namely very short term forecasting, short term forecasting, medium term forecasting, long term forecasting. Medium-term forecasting has a time range from a few weeks to the next 12 months. Medium-term electricity sales forecasting is usually used for planning and budget allocation for the next period [1] (Abraham et al., 2022).

This research proposes a model for forecasting electricity sales in the medium term, namely the next 6 months, using postpaid customer data and the LSTM, SARIMA forecasting method.

II. Related Works

A. *Previous Study*

Forecasting models have been carried out by many previous researchers. a lot of forecasting uses deep learning (LSTM) such as in Weather Prediction research with LSTM [3] (Lattifia et al., 2022). And there are also those who use SARIMA, such as in the research Forecasting the Semarang City Consumer Price Index Using SARIMA [2] (Dimashanti, 2021) or SARIMA Models for Predicting Inflation Levels [4] (Rizki and Taqiyyuddin, 2021) where each model is adjusted to the type of data or number of each study.

B. *SARIMA Models for Predict Inflation Levels*

This research conducted by [4] predicted the level of inflation occurring in Indonesia, using secondary data obtained from Bank Indonesia. This data occurred in the time period from January 2003 to November 2020. The modeling used involves all data without dividing it into train data or test data, with predictions for the next 7 months from December 2020 to June 2021.

This research uses SARIMA as a forecasting model. Based on the smallest AIC criterion, namely -238.10, the correct model formation result is $(1,0,1)(1,1,1)^{12}$. Then, based on the MAPE criteria, the SARIMA model $(1,0,1)(1,1,1)^{12}$ is obtained. Apart from that, a 95% confidence interval was also obtained for the predicted value obtained.

C. *Forecasting the Semarang City Consumer Price Index Using SARIMA*

This research conducted by [2] forecasts the Consumer Price Index (CPI), where the CPI is an important economic indicator that can provide information regarding developments in the prices of goods and services paid by consumers and is commonly used to measure the level of inflation. The CPI is also taken into consideration in regional workers' income. In the research, the data came from the Semarang City Central Statistics Agency (BPS) from January 2014 to December 2018, totaling 60 data.

From the estimated 170 models, parameter significance tests were carried out and then 9 temporary models were obtained. Of the 9 models, the best model was found to be SARIMA (1,1,1)(2,1,0)12 with an MSE value of 0.3639. From the SARIMA model, forecasting was carried out from January 2019 to December 2021. From this research it was found that the CPI value for the city of Semarang tends to remain unstable so that if this predicted value is used as a reference, a policy from the local city government is needed to stabilize the market so that it has an impact on CPI value.

D. *Weather Prediction Model Using the LSTM Method*

This research was conducted by [3] to create a model to predict weather that changes during the day and night. With the aim that if weather predictions are carried out accurately, it can improve human performance in activities. The train data is rainfall and temperature for 2013 – 2020 and the test data is for 2021.

Research was conducted on factors that influence weather, namely rainfall and temperature. From the rainfall data model with epoch 100, batch size 50 gives the smallest RMSE value, namely 1.7444, then the results of the rainfall and temperature tests show that the model with epoch 100, batch 50 gives the smallest RMSE and MAPE values. From the LSTM epoch 100 and batch 50 models that were formed to produce the appropriate output, the original temperature data pattern and the predicted temperature data pattern were not much different

III. PROPOSE METHOD

Based on the amount of data available, several model options perform forecasting. So in this research we will try to use 2 methods for forecasting, namely using LSTM and using SARIMA

A. *Long Short Term Memory (LSTM)*

Recurrent neural networks are a type of Deep Learning networks. Among the different types of RNN, Long ShortTerm Memory is very useful for time series data. RNN's can internally maintain the memory of the data that is input to them. Nevertheless, RNNs suffer from vanishing gradient problem which means that the model may not learn at all or learning becoming too slow. LSTMs were known to provide a solution to this problem. There are three gates in an LSTM. They are the input gate, forget gate and an output gate. These gates are activated on the sigmoid function which works on the range 0 to 1. The following figure shows the architecture of an LSTM.

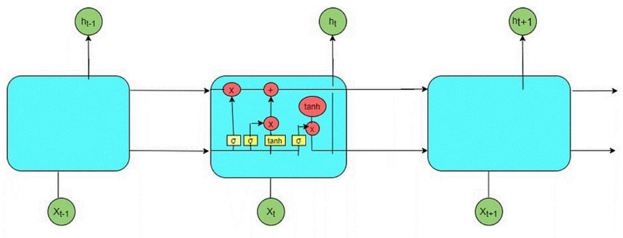


Figure 1 LSTM architecture

B. *Seasonal AutoRegressive Moving Average*

SARIMA model is the product of seasonal and non-seasonal polynomials and is designated by SARIMA (p, d, q) x (P, D, Q)S, where (p, d, q) and (P, D, Q) are non-seasonal and seasonal components, respectively with a seasonality's'. SARIMA model was defined ae equation 1

$$\Phi(B^S)\ \varphi(B)(1-B^S)^D\,(1-B)^d\,yt = \Theta(B^S)\ \theta(B)\varepsilon_t \qquad (1)$$

Where : Φ and φ = autoregressive (AR) parameters of seasonal and non-seasonal components, respectively; θ and Θ = moving average (MA) parameters of seasonal and non-seasonal components, respectively; B = backward operator, $B(y_t) = y_{t-1}$; $(1–B^s)^D = D^{th}$ seasonal difference of season s; $(1 – B)^d = d^{th}$ non-seasonal difference; ε_t = an independently distributed random variable; P and p = the orders of the AR components; Q and q = the orders of MA components; D and d are difference terms.

Four sequential steps as described below were followed for SARIMA modelling and forecasting.

- Model identification
- Parameter estimation
- Diagnostic checking
- Forecasting and performance evaluation of models

IV. DATASET AND METHODOLOGY

The dataset used is a postpaid customer dataset that is issued every month through the PLN application system or what is known as sorek. The dataset starts from January 2017 to December 2023, each month has its own dataset. In each dateset there are 73 features that will be preprocessed and processed

Table 1 73 features of dataset

THBLREK	UNITUP	RPANGSA	SLAWBP_PASANG	FAKM
IDPEL	TARIF	RPANGSB	SAHWBP	FAKMKVARH
NAMA	KDPT	RPANGSC	SLAKVARH	TGLCABUTPASANG
ALAMAT	KDPT_2	RPMAT	SAHKVARH_CABUT	DLPD
NOBANG	DAYA	RPPLN	SLAKVARH_PASANG	DLPD_LM
KETNOBANG	KDPROSESKLP	RPTAG	SAHKVARH	DLPD_FKM
RT	POSTINGBILLING	RPBK1	PEMKWH	DLPD_KVARH
NODLMRT	MSG	RPBK2	JAMNYALA	DLPD_3BLN
KETNODLMRT	RPPTL	RPBK3	PEMKVARH	DLPD_JNSMUTASI
RW	RPTB	SLALWBP	KELBKVARH	DLPD_TGLBACA
KODEPOS	RPPPN	SAHLWBP_CABUT	DAYAMAKS	ALASAN_KOREKSI
KDGARDU	RPBPJU	SLALWBP_PASANG	DAYAMAX_WBP	JAMNYALA600
NAMAGARDU	RPBPTRAFO	SAHLWBP	PEMDA	JAMNYALA400
KDDK	RPSEWATRAFO	SLAWBP	KOGOL	
UNITAP	RPSEWAKAP	SAHWBP_CABUT	SUBKOGOL	

A total of 84 monthly sore files were combined into one excel or CSV file to facilitate data processing. The next stage is preprocessing by doing:

Deleting features that are not needed in the research, so that the features used become the initial 4 features:

Table 2 features used in the research

THBLREK	IDPEL	RPTAG	PEMKWH

Then check the NaN value for the PEMKWH feature and fill all rows to 0. Add up the PEMKWH features in each file, which will then be combined into one file as follows:

Table 3 fixed dataset

BULAN	TAHUN	PEMKWH
12	2023	19102932
11	2023	19085002
10	2023	20051352
9	2023	16900918
8	2023	17623051
7	2023	17761832

A. *LSTM*

In the initial stage, the data is split into training data and test data with a ratio of 80: 20. Then a model is formed based on the existing data with several parameters used.

Table 4 LSTM Model Propose

	model 1	model 2	model 3	model 4	model 5
lstm unit 1	50	50	100	50	100
lstm unit 2	50	50	50	50	50
dense 1	25	25	25	25	25
dense 2	1	1	1	1	1
optimizer	adam	adam	adam	adam	adam
batch size	32	1	1	1	1
epoch	50	100	100	50	50

In this research, several scenarios were carried out with the consideration that when the initial testing was carried out, the MAPE and RMSE values obtained were deemed to be not good, and besides that, the amount of data tended to be small. In the tests carried out, it was found that the best matrix value was in model 2, namely RMSE 114.78 and MAPE 9.29.

Then the predictions for the future period are tested, which are compared with the actual data using model number 2

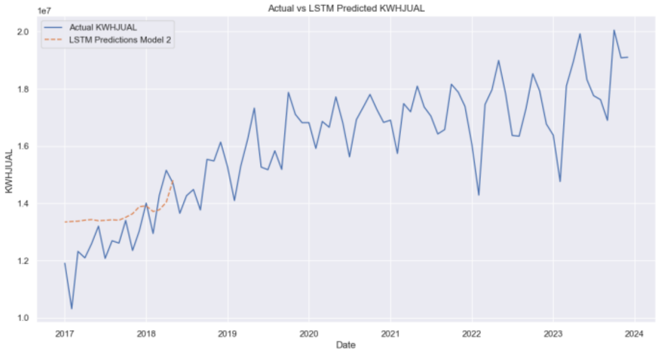


Figure 2 LSTM Prediction

B. *SARIMA*

In the SARIMA model, the ACF and PACF plots are carried out first

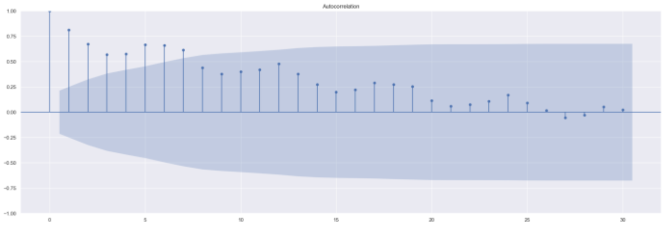


Figure 3 ACF Plot

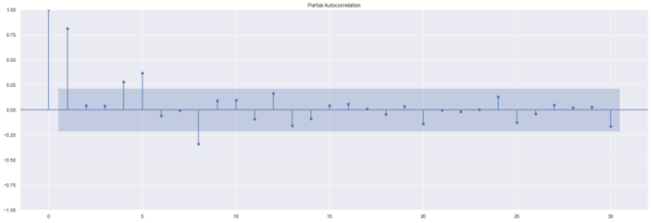


Figure 4 PACF Plot

Based on the plot above, it can be concluded that the pattern formed can be predicted by several values, namely in the ACF plot the non-seasonal MA order can be identified looking at the significant lag, so we get MA(0) and MA(1) because of the cut off after the 1st lag, then for the seasonal MA orders MA(0) and MA(1) because the cut off is at the 28th lag or the 1st seasonal lag. On the PACF plot the non-seasonal AR orders AR(0), and AR(1) can be identified because at PACF plot cut off at the 1st lag. Then for seasonal AR orders AR(0), and AR(1) because in the PACF plot the cut off is at the 7th lag. Meanwhile, the order for differencing is d=0 for non-seasonal and D=1 for seasonal because seasonal differencing is carried out once, so there are several possible SARIMA models that can be formed.

Then parameter estimation was carried out and the best model was selected, from several test results the selected SARIMA model with the smallest MAPE was SARIMA(1,1,1)(1,1,1)^{12}. From this model, it is used to make predictions with the following graph:

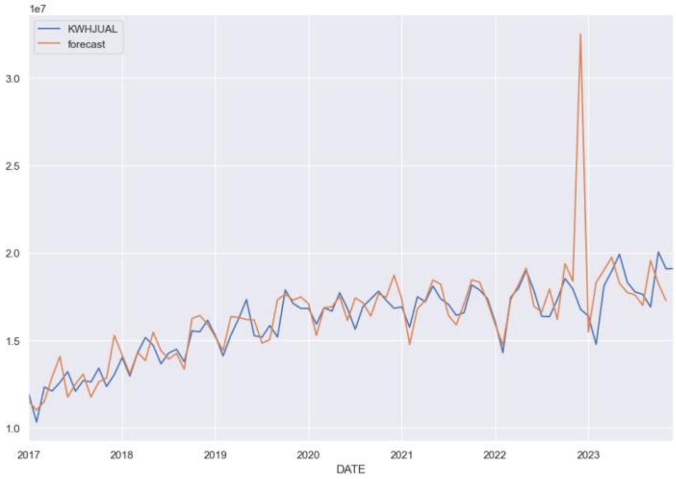
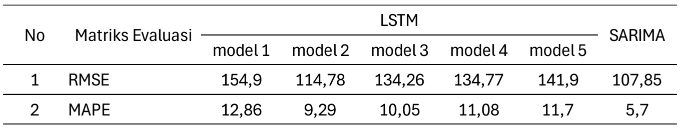


Figure 5 Actual vs Predicted SARIMA Model

From the research results, SARIMA has a better error value compared to LSTM, here are the details:

Table 5 Error Metrics Values



No	Matriks Evaluasi	LSTM					SARIMA
		model 1	model 2	model 3	model 4	model 5	
1	RMSE	154,9	114,78	134,26	134,77	141,9	107,85
2	MAPE	12,86	9,29	10,05	11,08	11,7	5,7

From these results, overall SARIMA gives a better value, this happens because basically LSTM is used to carry out forecasting with large amounts of data, whereas in this study, although initially the numbers were large, after preprocessing they became smaller. Meanwhile, SARIMA is better at forecasting, which in this research will show seasonal forecasting.

V. CONCLUSION

In this research, it can be seen that the amount of data affects model performance, where LSTM requires a fairly large amount of data. For the next step, if the amount of data has not increased, other efforts need to be made to improve the performance of the model, such as combining it with other machine learning or using other techniques. while for SARIMA, existing data can already be used. In general, the performance of the SARIMA model is better in predicting electricity sales compared to LSTM where the RMSE value is 107.85 and MAPE is 5.7

REFERENCES

[1] Abraham, R., Samad, M., Bakhach, A., El-Chaarani, H., Sardouk, A., Nemar, S., and Jaber, D. (2022): Forecasting a Stock Trend Using Genetic Algorithm and Random Forest, *Journal of Risk and Financial Management*, **15**(5), 188. https://doi.org/10.3390/jrfm15050188

[2] Dimashanti, A. R. (2021): Peramalan Indeks Harga Konsumen Kota Semarang Menggunakan SARIMA Berbantuan Software Minitab, **4**.

[3] Lattifia, T., Buana, P. W., and Rusjayanthi, N. K. D. (2022): Model Prediksi Cuaca Menggunakan Metode LSTM, *JITTER : Jurnal Ilmiah Teknologi dan Komputer*, **3**(1), 994. https://doi.org/10.24843/JTRTI.2022.v03.i01.p35

[4] Rizki, M. I., and Taqiyyuddin, T. A. (2021): Penerapan Model SARIMA untuk Memprediksi Tingkat Inflasi di Indonesia, *Jurnal Sains Matematika dan Statistika*, **7**(2). https://doi.org/10.24014/jsms.v7i2.13168

PERNYATAAN

Dengan ini saya menyatakan bahwa makalah yang saya tulis ini adalah tulisan saya sendiri, bukan saduran, atau terjemahan dari makalah orang lain, dan bukan plagiasi.

Bandung, 12 Juni 2024
Ttd
Paulus Saritosa - 23522315