Data Mining Implementation to Predict Sales using Time Series Method

Agung Triayudi
Information Communication Technology Department
Universitas Nasional
Jakarta, Indonesia
agunatriayudi@citivitas.unas.ac.id

Thoha nurhadiyan H
Information Communication Technology Department
Universitas Serang Raya
Banten, Indonesia
thoha.nurhadiyan@gmail.com

Sumiati
Information Communication Technology Department
Universitas Serang Raya
Banten, Indonesia
sumati82@yahoo.com

Vidila Rosalina
Information Communication Technology Department
Universitas Serang Raya
vidila.suhendarshah@gmail.com

Abstract—Sales transaction data histories can be used to predict the possibility of sales transaction that will occur in the future. These characteristics are in accordance with forecasting using time series method where this method uses previous data as tools to predict transaction value that will appear in the present time. Company X that runs its business by sell their product through distributors has sales data that is not optimally utilized. The average number of sales per year ranges from 5000 transactions which is not use to forecast transactions hereafter. Transaction data is stored in the company database so that data mining technology can be applied to support company X transaction data collection from previous year. The data is processed in applications where the results of forecasting are compared with real data in 2018 to see the accuracy of the forecasting results. The graphic that shown in application has pattern which can use for forecasting. From the forecasting method used, it can be seen that the forecasting results show data that came out did not produce data that matched the real data where the highest level of accuracy was 99.68% and the lowest accuracy was still above 50%.

Keywords—sales transaction data, data mining, forecasting, time series, accuracy

I. INTRODUCTION

Transaction is the core activity from a business process and a cycle that consistently done in several period of time. Transaction started from the very beginning of a business process began its activity. Company will record and documented every transaction in business process as a cycle in financial management [1]. There are several advantage and activity that can be done from transaction documentation, one of which is to do forecasting from transactions of a business process to be a part of decision support making from the business activities carried out.

Forecasting is a part from planning a business process that includes many things such as financial management, investment decision making, operational budgeting, until producing and introducing a new product [2]. Forecasting become an important tool to predict an accurate transaction process with the result, so that business process can work on efficiently.

Company X is a company that sells its product through distributors with selling transactions data recorded in their database. Monthly transactional data in company X could achieve number between 300 transactions with value above 400 million Rupiah. With that amount of data it is still not utilized by the company to the fullest. As a sample case that will be discussed in this journal, the company made forecasting manually, that is by exporting transaction data from database to Ms. Excel format and then they will study the data from Ms. Excel one by one.

The authors proposes to utilize available data on database in company X to be processed automatically by application that specialized for forecasting sales transaction.

The authors will also utilize data mining that is used to quarry [3] or retrieve all data that has been recorded in the company’s database. The process in data mining has the same characteristic as forecasting in this study using the time series method where both of them using data that already exist. The output of processed data will be obtained a pattern of data sales transaction from the company [4].

II. RESEARCH METHOD

The research process is described using flowcharts shown in Figure 1 to display the research steps.

A. Datasource Research

Data resource for this research comes from sales transaction of company X that sells its product through distributors in Indonesia. The data taken from January 2015 until December 2018 with range around 5,000 data per year. The data has been masked for confidentiality reason.

B. Forecasting Time Series

Forecasting procedures are estimating future values based on patterns in data set [3], [4]. Looking for the case that we discussed related to the value of future transaction based on the value of previous transactions. Time series is one of method in forecasting. Time series can be interpreted as collection from successive observations [4], [5]. The main components that affect forecasting values are [6]:

1. Trend
2. Seasonal
3. Cyclical
4. Random
C. Database and Application

Database is a media for storing the data that will utilize to process existing data. Database created based on modelling of Class Diagram which is core of Unified Modeling Language (UML). Class diagram can represent the tables of database with the attributes and also the relation as foundation for creating the forecasting application.

Use Case Diagram is a part of UML that will used to portray the elation of users with the features in application. The two of the UML models create to facilitate for developing the application.

D. Moving Average

Forecasting of a time series has several methods, one of the methods is moving average. The moving average method works by eliminating or reducing random data (randomness) in a time series. This method can be achieved by averaging several similar data values at once, where positive and negative errors that might occur are removed or omitted [7], [24]. Moving average forecasting utilize some actual data in the previous time to produce forecasting [6], [8]. The formula for getting the moving average value for a period of a time is as follows.

\[ F_{t+1} = \frac{X_t + X_{t-1} + \cdots + X_{t-(n+1)}}{n} \]  

\( F_{t+1} \): forecast value for the time period to \( t+1 \)  
\( X_t \): actual value for \( t \) period  
\( n \): amount of data

E. Forecasting Evaluation

Forecasting evaluation has several formula to help determine forecasting value, these formula are mean absolute deviation (MAD), mean squared error (MSE), and mean absolute percentage error (MAPE) [2], [9]. Mean absolute deviation or MAD is a common method to get to know mistakes or irregularities from forecasting with formula as follows:

\[ MAD = \frac{1}{n} \sum_{i=1}^{n} |Actual - Forecast| \]  

Mean squared error or MSE functioned to compare forecasting data with the data source to minimize errors from prediction results. The equation of formula is as follows:

\[ MSE = \frac{\sum_{i=1}^{n} (Actual - Forecast)^2}{n} \]  

Mean absolute percentage error or MAPE with minimum results is the best prediction result [2]. When theres no strange data or zero valued data. With formula:

\[ MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{Actual - Forecast}{Actual} \right| \times 100\% \]

N \( : \) AMOUNT OF PERIOD

F. Data Mining

Data mining can be defined as search for relationship and global patterns that are in a large and hidden database because of the size of a database [10]. Furthermore, data mining also functioned to predict trend and behavior in the future, and allows business to make knowledge based decisions [11], [19].

To get big data prediction, data mining defined into four main functions [12], [13]. The function shown in Figure 2.

\[ X_t + X_{t-1} + \cdots + X_{t-(n+1)} \]
defined as the process of separating labeled data sets into groups based on similarity [17], [25].

- **Classification**
  Data classification is processed based on mining processed needs.

- **Prediction**
  It’s a core process where data taken and then processed with method or certain formula to do forecasting in the future. Time series method is used for doing these process [21], [22].

- **Association**
  In this stage, association will be replaced by doing an evaluation from the result of forecasting. Evaluation will use MAD, MSE, and MAPE method as measurement tools to determine accuracy from forecasting results [14], [23].

III. RESULTS AND ANALYSIS

The designing process of an application started from database planning where database’s fields arranged based on collected data. As mentioned in methodology, application design using UML as tools with class diagram used as tools that help to represent database structure that will be built by showing the class of the application and its attributes [18], [20]. Figure 3 illustrates the relation from the transaction table that represent class from transaction as process core of an application. Transaction table has relation with table tdoc, mdistributor, and mproductstock. Product data taken from the productstock which has data on the number of products available when transaction is made.

Based on Figure 3, database built by using PostgreSQL which is a relational database management system. PostgreSQL or postgres selected because it can be obtained for free, open source, light, and it’s built with java programming language which is an object-oriented programming language. As for the Postgres version used is 9.5. The application built has several users with different privilege for each user. There are two main users for the application, they are the operator and the forecaster. Operator has several function like transaction data input, while forecaster is the main user for the application where forecaster has privilege to see the result of forecasting based on transaction that recorded in application. For more details the use case diagram in Figure 4 will illustrates the privilege of each user in the application.
Furthermore, the application will be built using Java programming language with Java runtime environment version 1.8. The application also using framework zkoss or known as ZK which a framework for web-based java enterprise.

A. Implementation and Application Testing

- **Clustering**
  The collected data are transformed into tables in database that have been designed according to the class diagram. These data then taken (mining process) from database which will grouped based on year and month.

- **Classification**
  The amount of data collected is around 21,305 transaction of data. Of this amount there are data in 2014 which were recorded only in December, that makes data like this cannot be used as a source of data for forecasting. Data displayed from 2015 to 2017 and grouped by month. The results of the grouping will used as data sources for forecasting every month. For more details can be seen in Figure 5 where data is displayed in January 2015 to December 2017.

- **Prediction**
  Table 1. Is a data source that forms a graph in Figure 5 that will be used for forecasting. Can be seen from the graph, it is shown in Figure 5 that transaction data has the same pattern so that the data can be forecasted using a time series model. In this study, forecasting will produce data for 2018. The results of the forecasting will be compared with real data in 2018 to assess the accuracy of forecasting using applications that have been built.

<table>
<thead>
<tr>
<th>Month</th>
<th>2015</th>
<th>2016</th>
<th>2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan</td>
<td>516,159,200</td>
<td>540,066,500</td>
<td>539,130,800</td>
</tr>
<tr>
<td>Feb</td>
<td>503,928,600</td>
<td>532,178,600</td>
<td>529,788,700</td>
</tr>
<tr>
<td>Mar</td>
<td>498,272,000</td>
<td>567,840,500</td>
<td>559,943,400</td>
</tr>
<tr>
<td>Apr</td>
<td>447,098,800</td>
<td>553,698,000</td>
<td>670,138,600</td>
</tr>
<tr>
<td>Mei</td>
<td>452,022,800</td>
<td>564,036,000</td>
<td>930,170,600</td>
</tr>
<tr>
<td>Jun</td>
<td>1,210,802,200</td>
<td>1,163,893,700</td>
<td>1,166,818,100</td>
</tr>
<tr>
<td>Jul</td>
<td>702,191,200</td>
<td>568,383,700</td>
<td>667,855,700</td>
</tr>
<tr>
<td>Aug</td>
<td>459,480,100</td>
<td>592,436,400</td>
<td>616,369,500</td>
</tr>
<tr>
<td>Sep</td>
<td>476,830,700</td>
<td>621,444,600</td>
<td>643,756,800</td>
</tr>
<tr>
<td>Oct</td>
<td>428,345,300</td>
<td>599,556,300</td>
<td>643,373,100</td>
</tr>
<tr>
<td>Nov</td>
<td>471,286,400</td>
<td>595,332,500</td>
<td>660,943,400</td>
</tr>
<tr>
<td>Des</td>
<td>908,760,100</td>
<td>908,058,500</td>
<td>842,217,900</td>
</tr>
</tbody>
</table>

Figure 6 shows the results of the forecasting carried out with the application. The results show that real transactions in 2018 (orange lines) tend to have higher numbers compared to data from forecasting results indicated by purple lines. In January, June and December the numbers tend to be the same.

![Graph of Transaction and Forecast](image1)

Fig. 5. Data graphic Jan 2015 until Dec 2017.

![Graph of Transaction and Forecast](image2)

Fig. 6. Forecasting results.
From pattern formed that shown in Figure 5. There are irregular patterns in May, June and December. However, the results in Figure 6 show that in June and December it should be difficult to produce numbers to predict but instead produce numbers that tend to be the same. For more details the results data are presented in the form of a Table 2.

<table>
<thead>
<tr>
<th>TABLE II. FORECASTING RESULT DATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real 2018</td>
</tr>
<tr>
<td>Jan</td>
</tr>
<tr>
<td>Feb</td>
</tr>
<tr>
<td>Mar</td>
</tr>
<tr>
<td>Apr</td>
</tr>
<tr>
<td>May</td>
</tr>
<tr>
<td>Jun</td>
</tr>
<tr>
<td>Jul</td>
</tr>
<tr>
<td>Aug</td>
</tr>
<tr>
<td>Sep</td>
</tr>
<tr>
<td>Oct</td>
</tr>
<tr>
<td>Nov</td>
</tr>
<tr>
<td>Dec</td>
</tr>
</tbody>
</table>

From the results of forecasting it can also be seen that the results of forecasting data are average of all data sources. Though the latest data should have a greater value than the source data [17].

- **Association**

  A forecast always contains an element of uncertainty. This element of uncertainty causes errors or deviations in generating forecasting values [17]. The MAD, MSE, and MAPE methods help measure the magnitude of deviations from forecasting results.

From the Table 3 shows the amount of deviation that occurs in the forecasting process. The results of the MAPE method are easier to see, such as in December which has a value of 0.32%. This means that the forecasting accuracy in December reached 99.68%. In sequence, December (99.68%), June (98.97%), and January (98.26%) are forecasting results with a high degree of accuracy. Forecasting results and error measurements on the forecasting results with a high degree of accuracy. This is because from forecasting results.

<table>
<thead>
<tr>
<th>Table 3. With MAD values 9,378,800, MSE 87,961,889, and MAPE 1.73.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAD</td>
</tr>
<tr>
<td>Jan</td>
</tr>
<tr>
<td>Feb</td>
</tr>
<tr>
<td>Mar</td>
</tr>
<tr>
<td>Apr</td>
</tr>
<tr>
<td>May</td>
</tr>
<tr>
<td>Jun</td>
</tr>
<tr>
<td>Jul</td>
</tr>
<tr>
<td>Aug</td>
</tr>
<tr>
<td>Sep</td>
</tr>
<tr>
<td>Oct</td>
</tr>
<tr>
<td>Nov</td>
</tr>
<tr>
<td>Dec</td>
</tr>
</tbody>
</table>

In the picture it can be seen that the calculation using the application is the same as the calculation manually. Forecasting data can be seen in Table 2. In January with a value of 531,785,500 with the results of error calculations in January in Table 3. With MAD values 9,378,800, MSE 87,961,889, and MAPE 1.73.

**Fig. 7. Screenshot of manual calculation.**

**IV. CONCLUSION**

There are several points that can be taken from this study:

- Forecasting is almost impossible to produce data accuracy of 100%. In this case, the highest forecasting value with the calculation of MAPE is 99.68%, namely in December. The lowest yield forecasting occurred in May with forecasting accuracy still above 50%. Even so, forecasting can still be used to bring up numerical estimates that can be used in business processes assisted by the patterns formed on the graph.
- In the time series method, Trend (T) and Season (S) greatly affect data movement. In this study, the trend pattern that occurs actually provides forecasting accuracy that is close to accurate. This can be seen in June and December where the graphics soared, but had a forecast accuracy of more than 98%.
- Variable increase in value per year ignored by forecasting with time series methods can also be a cause of reduced forecasting accuracy. This is because the formula of the time series method is the average of the source data without added value. Table 2 shows this, where the results of forecasting are not higher than real data.
- Applications are very possible to be developed at the level of forecasting and development for business processes. Trend and Season are variables that are very influential on the movement of data and can also be formulated and made parameters to help improve forecast accuracy.

**ACKNOWLEDGMENT**

This research is the result of the basic research scheme of the Indonesian DIKTI grant B/877/E3/RA.00/2020.

**REFERENCES**


