Sales prediction of Four Wheelers Unit (4W) with seasonal algorithm Trend Decomposition with Loess (STL) in PT. Astra International, Tbk.

Abdi Telaga¹, Anggun Fadhlin Librianti¹ and Umairoh Umairoh¹
¹Astra Manufacturing Polytechnic, Indonesia

E-mail: abdi.telaga@polman.astra.ac.id

Abstract. Corporate Information System and Technology (CIS&T) division in PT Astra International Tbk is digitizing all business lines. Therefore, the company started to utilize machine learning to predict four wheelers (4W) sales in 2017. Data used were car sales data from 2014 to 2016. The data were obtained from company’s data warehouse (DW). Further, to develop application, firstly, Minitab was used to compare algorithm in modelling phase. Then, Azure Machine Learning Studio (Azure ML) was utilized to build prediction model based on algorithm chose in first phase. Finally, visualization was developed using Power Business Intelligence (Power BI). Time series algorithm used to solve the case is Seasonal Trend Decomposition with Loess (STL). While, machine learning model uses R library, and process flow module in Azure ML. The STL prediction model yield 9.38% error. Visualization of result is divided into two sections filtered by three parameters type, color and area. The first section in dashboard is sales data report showing percentage of type, color of cars and deployment of 4W units in Indonesia through ArcGIS Maps. The second section is forecast result containing chart of x axis to represent year and y axis to represent number of units. Further, the chart provide historical, current and forecast information. The research contributes to illustrate development of prediction using machine learning in industrial environment.

1. Introduction

1.1. Background

Increasing of market demand for vehicle products gives a great opportunity to PT Astra International, Tbk. (Astra) as the largest automotive brand in Indonesia. As a market leader of automotive, Astra demanded to innovate continuously. In order to balancing company with its technology, Astra digitizes all business lines. Implement machine learning to maximize data functions in Astra's data warehouse is one of the ways.

In distribution process of vehicles to various branches, there is often a lack of space in the main dealer. It will cause dealer lease new space that will incur additional costs. In addition, unit prices will go down drastically because main dealers must provide discounts or price reductions to attract consumers. Main factor of the occurrence of slow moving and deadstock is the difference in the number and specifications sent to the main dealer needs.

Prediction of production demand for products and services is very important in planning and controlling production. Besides being able to predict future, it is necessary for stakeholder to make plans.
Data sample used in this research is 4W Toyota Sales Operation (TSO) unit data. Decision to use TSO data as a sample is because Toyota Astra Motor has the largest percentage of car sales in Indonesia, which is around 36.5% in 2016 (TEMPO.CO, 02 January 2017). Machine learning will process data from TSO to creating model and predicting sales in 2017.

1.2. Goals
The purpose of this research is to predict sales of 4W units in 2017 with parameters using are type, color and area. It aims to be able to adjust the units distributed and needed for each dealer.

1.3. Limitations
To create 4W unit prediction model, the limitations are:
1. Raw data is sales data of 4W unit TSO January 2014 - December 2016.
2. Research uses data marts that formed from data warehouses with type, color, and area parameters.
3. Modelling using Microsoft Azure Machine Learning (Azure ML) Studio program and visualization using Power Business Intelligence (Power BI).

2. Study Literature

2.1. Phases
Cross-Industry Standard Process for Data Mining (CRISP-DM) is a standard developed in 1996 which is intended to conduct an analysis process of an industry as a problem-solving strategy of a business or research unit (Peter, 2012). There are no specific conditions or characteristics for data to be processed by CRISP-DM, because it will be processed again in the phases in it. The CRISP-DM phase can be seen in Figure 1.

![Figure 1. CRISP-DM Phases (Chapman, 2000)](image)

There are six phases in CRISP-DM, namely: (Data Mining Methods and Models, 2006)

A. Business Understanding Phase
1. Determination detail of project objectives and needs in business scope or research unit as a whole.
2. Translating goals and boundaries into a formula for data mining problems.
3. Prepare an initial strategy to achieve goals.

B. Data Understanding Phase
1. Collect data, if data source more than one database then data integration process or data integration will be carried out.
2. Develop data inquiry analysis to further identify data and search for initial knowledge.
3. Evaluating data quality, checking data and cleaning invalid data or data cleaning processes.
4. If desired, select a small group of data that might contain patterns of problems.

C. Data Preparation
1. Prepare initial data, data collection that will be used for entire next phase or data selection process.
2. Select case and variable to be analyzed, according to the analysis that will be carried out.
3. Make changes to variables if needed.
4. Prepare initial data so that it is ready for modelling or data transformation.

D. Modelling
1. Select and apply appropriate modelling techniques.
2. Can use several of same techniques for the same problem.
3. Can return to data processing phase if needed to make the data into a specific form.

E. Evaluation
1. Evaluate one or more models used in the modelling phase or the Evaluation Pattern process.
2. Determine whether the model is in accordance with the objectives in the initial phase.
3. Are there important issues of business or research that are not handled properly.
4. Making decisions related to the use of results from data mining.

2.2. Time Forecasting Series
Time series are observations series of variables that will be observed sequentially from time to time and recorded according to sequence of events (Wei, 2006). Time series analysis is a prediction method based on past values or data from variable and past errors. The purpose is to find data patterns and extrapolate pattern to future periods.

2.3. Decomposition Time Series
Prediction model analyzes time series by processing past data into components and then projecting it forward. There are four elements of the time series that are decomposed by this model, namely: trends, seasons, cycles, and random variations.

3. System Analyzing

3.1. Business Understanding
This research begins with conducting business observations and observing data in Astra's data warehouse. Data observation was conducted through interviews with Business Analysts and Data Engineers. This interview aims to find out the business processes and data descriptions that can help research process. Based on the results of interviews, a special business flow and data mart was created in handling slow moving problems and four wheeler (4W) deadstock units at Astra.

As explained above, machine learning implementation in this research is to deal with slow moving and deadstock problems. This research is pursued by factors that do not match the number and specifications of the units distributed with the required main dealer in an area. Data mart will be made from Astra's data warehouse. The purpose of implementing machine learning in this research is to get sales prediction results that can be used to support the decision of the number and specifications of units sent to main dealer.

3.2. Data Understanding
Data understanding is the stage of collecting initial data and analyzing data to understand what can be done in those data. Data understanding refers to the Astra's data warehouse. Data collection is first carried out by retrieving the data contained in Astra's data warehouse. In this case, the data is processed
into a data mart so that the data used in case is only one table. After that, data quality evaluation will be carried out.

3.3. Data Preparation
Data preparation covers all dataset making activities that will be entered into a modelling tool from raw data or create a new datasheet of machine learning modelling. Data mart is independent learning or separate from operational database. Its main function is specifically for forecasting time series modelling tools. Data preparation is a stage to build new database as the end for modelling.

3.4. Dataset Description
The dataset design must adjust the formulation of business and data understanding. Tables design for machine learning process can be seen in table 1.

| No | Field      | Type   | Description          |
|----|------------|--------|----------------------|
| 1  | EquipmentNo| Varchar| Vehicle machine number |
| 2  | Color      | Varchar| Unit colour          |
| 3  | Price      | Number | Unit price           |
| 4  | Type       | Varchar| Vehicle unit type    |
| 5  | Area       | Varchar| Sales area           |

3.5. Data Selection
Stage of selecting data used for analysis is based on several criteria, its relevance to modelling purpose, quality and technical constraints such as limits on volume or data type. Data selected in each database is related to the need for predicting sales of 4W units.

4. System Design

4.1. Modelling
Algorithm used in this study is Seasonal Trend Decomposition with Loess (STL) Time Series Forecasting. STL helps decompose time series into 3 components, those are trend, seasonal and remainder. STL algorithm is chosen because match for the problems faced by predicting sales of 4W units. Creating model using Microsoft Azure ML Studio.

4.2. Evaluation
Azure ML Studio will use data to train model. After that, model testing is used to see how close prediction result with original data. The data can be seen in Figure 2.
Figure 2. Evaluation Data

Forecast period: forecast period is month description to predict data. Start from data in August, the 32th data with forecast period description 1.

Forecasting: forecasting is a data prediction result that resulted from model. For example, forecasting period 1 with forecast 30008.486. The data has a meaning that data in August 2016 will be predicted 30008.486.

Time: time is month description to actual data start from August, the 32th data.

Data: data is actual data start from August, the 32th data.

Description for the first row: Forecast period : 1
Forecast : 30008.486
Time : 32
Data : 35027

It means prediction in August 2016 is 30008.486 with actual data 35027. Data in rows 6-27 are prediction for January – December 2017. Validation testing tests how many errors made from prediction data against original data. Formula to calculate errors is:

$$\text{error} = \frac{|\text{prediction value} - \text{original data value}|}{\text{original data value}}$$

From the formula above, evaluation model result in this research can be seen in Table 2.

Table 2. Evaluation Model Result

| Month           | Real Data | Prediction Result | Error       | Error Percentage |
|-----------------|-----------|-------------------|-------------|------------------|
| August 2016     | 35027     | 30008             | 0.14328946  | 14.33%           |
| September 2016  | 34299     | 30750             | 0.1034724   | 10.35%           |
| October 2016    | 35279     | 33092             | 0.06199155  | 6.20%            |
| November 2016   | 36392     | 33324             | 0.08430424  | 8.43%            |
| December 2016   | 35140     | 32476             | 0.07581104  | 7.58%            |
| Rata-rata Error |           |                   | 0.093773738 | 9.38%            |
Predictive result data errors shows fluctuating data. Average of errors result is 0.09377 or 9.38%.

5. Result and Testing

In deployment process, visualization result of TSO 4W unit prediction model was shown by Power Business Intelligence application. Visualization consists of 2 parts and each has 3 parameters namely type, colour and area. First part is dashboard reporting extracted from sales data from 2014 to 2016. Analyzing process include showing type with car colour which is illustrated based on the percentage of sales and depicted deployment of Astra 4W units in Indonesia with ArcGIS Maps. Second part is forecasting results of machine learning model described by line charts. Axis x describes year and y represents unit number. Data is divided into 3 parts, those are historical, current and forecast. Before visualizing, prepare data from predicted 4W TSO unit to be input as visualization input. Predictive data is re-inputted to spreadsheet sheet that has combination unit available by adding several columns for prediction results. Visualization interface can be shown in Figure 3 and 4.

6. Conclusions and Recommendations

6.1. Conclusions

Based on 4W TSO forecasting testing model making, it can be concluded that:

1. Model and visualization of 4W unit sales prediction for 2017 were successfully completed using TSO sales data with type, colour and area parameters and STL algorithm.
2. The prediction model has an average error of 9.38%.
3. Each data has its own character that must be treated specifically with different data cleansing treatments.
6.2. Recommendations
Based on conclusions and process of making TSO 4W forecasting, suggestions that can be proposed are:

1. Implementing model for other sales operations, so that all sales operations have sales predictions.
2. Creating program by integrating results of Minitab, Azure ML Studio and Power BI in making 4W TSO forecasting, so that users can access all results with one application.
3. In visualization parameter, create check boxes so that users can choose more than one type of each parameter.

References
[1] T Larose D 2006 Data Mining Methods and Models. Hoboken (New Jersey: John Wiley & Sons Inc)
[2] Chapman P 2000 CRISP-DM 1.0 Step by Step Data Mining Guide SPSS Inc
[3] Gartner Magic Quadrant for Business Intelligence and Analytics Platforms gartner.com
[4] Kurniawan A 2016 Mengenal Microsoft Azure ML (Jakarta: PT Elex Media Komputindo)
[5] Mai Nguyen L F https://www.coursera.org/learn/big-data-machine-learning
[6] Microsoft Embed the wow of Power BI in your applications with Microsoft Power BI Embedded! powerbi.microsoft.com.
[7] Microsoft What is Power BI https://powerbi.microsoft.com/en-us/what-is-power-bi/
[8] Murphy K 2012 Machine Learning: A Probabilistic Perspective (London, England: The MIT Press)
[9] Peter H 2012 Machine Learning in Action (New York: Manning Publication)
[10] S J Russell P N 1995 Artificial Intelligence: A Modern Approach vol. 2 Prentice hall Englewood Cliffs