Using Machine Learning in Sales Predication

Yuxin Liu, Song Chen*, Mao Zheng
Department of Computer Science
*Department of Mathematics and Statistics
University of Wisconsin-La Crosse
La Crosse WI, 54601
mzheng@uwlax.edu

Abstract

Machine learning is a method of data analysis that automates analytical model building. We are interested in exploring machine learning algorithms to study past sales data in order to make future sales predications for Fastenal Company.

Fastenal Company is the largest fastener distributor in North America. It has over 2,600 branches throughout the US, Canada, Mexico and Europe along with 13 distribution centers. Optimizing the inventory in the distribution centers and branches will certainly reduce shipping and storage costs for the company. In order to do so, we are using Fastenal's product consumption table from the period of 2013 to 2016 to predict product sales in the future as a starting point. In our approach, we used Linear Regression, Autoregressive Integrated Moving Average (ARIMA) and Long Short Term Memory (LSTM) to forecast product sales in the selected branch or distribution center. To evaluate the forecast result, we used the Coefficient of Determination (R^2 for short), and Root Mean Squared Error (RMSE) to measure the accuracy of our model.

The contribution of this research work is that we define a process flow to use the three models in practice. We ran the data through a Linear Regression model first. Based on the model accuracy measurement, we either accept the forecast result or pass the bad fit data into the ARIMA model. Similarly, we then either accept the ARIMA model result or pass the bad fit data to the LSTM model. This process is proposed due to the limitation of computation power. The size of the consumption table is over 146 millions lines, and the three models have different processing times. The Linear Regression is the fastest, and the LSTM is the slowest.

The future work of this research includes improving model accuracy and plan inventory redistribution based on the forecasted sales, taking into consideration the shipping distance and the cost.

1 Introduction

Fastenal Company is an American company based in Winona, Minnesota. Through distributing goods used by other businesses, it is the largest fastener distributor in North America. It has over 2,600 branches throughout the US, Canada, Mexico and Europe along with 13 distribution centers. These regional distribution centers distribute products to either wholesale customers or branches. Currently some branches have a surplus of items while other do not have enough. Optimizing the inventory in the distribution centers and branches will certainly increase the sales and reduce shipping and storage costs for the company. Fastenal Company hopes to automate the distribution center manager's manual sales predication work and to improve the accuracy of sales predication results. Hence, we propose using machine learning to automate this work.

Machine learning is a field of computer science that gives computer systems the ability to "learn" with data, without being explicitly programmed [1]. In this project, we are interested in exploring machine learning algorithms to study past sales data in order to make sales predications for Fastenal Company. We have been given Fastenal's product consumption table from the period of 2013 to 2016**.** We used the Linear Regression, Autoregressive Integrated Moving Average (ARIMA) and Long Short Term Memory (LSTM) to forecast product sales in the selected branch or distribution center. To evaluate the forecast result, we are using Coefficient of Determination (R^2 for short), and Root Mean Squared Error (RMSE) to measure the accuracy of our model.

We developed a web-based application for sales predications. The manager can select a regional distribution center, select a branch that belongs to that distribution center, then select a particular item and choose one of the three machine learning algorithms to forecast sales. The application can also display the predication data from all three algorithms and the two accuracy evaluations.

This web-based application uses Python as the programming language, Django framework and MySQL database in the backend.

2 Machine Learning Algorithms

In this section, we briefly introduce three machine learning algorithms used in our project.

2.1 Linear Regression

Linear Regression is a mathematical model that allow us to draw a straight line from the given dataset, and predict the future based on the extrapolation of the line. A linear regression line has the equation of the form $\boldsymbol{Y = a + bX}$, where $\boldsymbol{X}$ is the explanatory

variable and Y is the dependent variable. The slope of the line is b, and a is the intercept (the value of y when $x = 0$). In our project, we use every month as the explanatory variable, and the monthly sale's quantity of each item as the dependent variable. Based on Fastenal's four year consumption table, we generated a linear regression model for each item. Following the line, we are able to predict the sales for the future months. However, following the straight line, the forecasted sale will be either increased or decreased value. The linear regression model cannot reflect the factor of seasonal changes for the product demand.

2.2 Autoregressive Integrated Moving Average (ARIMA)

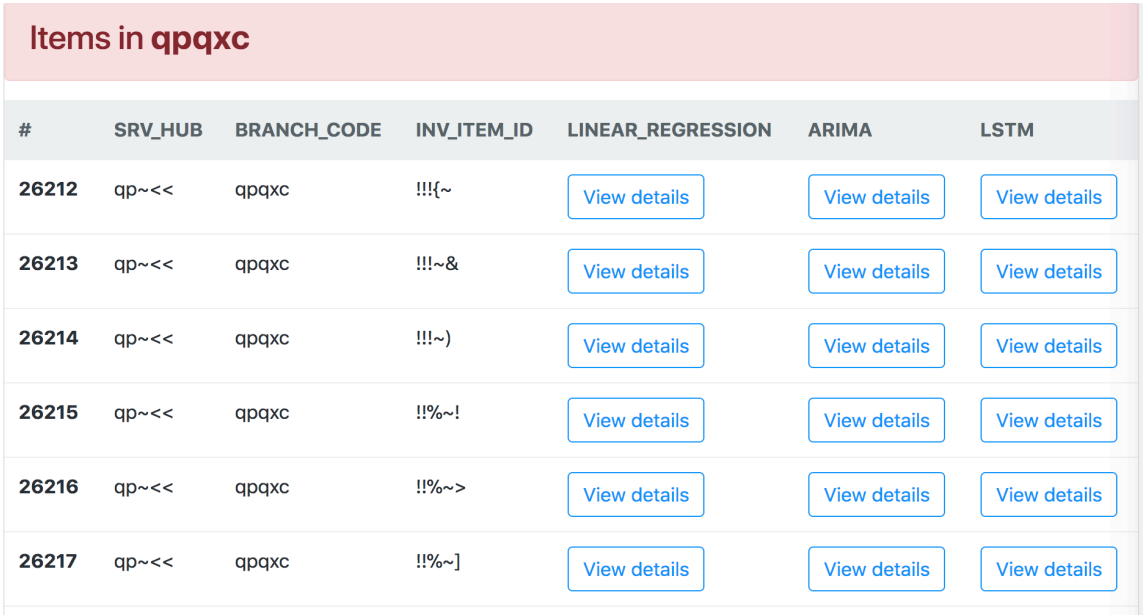
Fastenal's four consumption table is the sales data for every item per month. Hence time plays a very important role in the given data set. In order to explore seasonal changes for the sale's quantity, we considered applying time series models. A time series is a series of data points indexed in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time [2]. By analyzing time series data, we can extract meaningful statistics and other characteristics. It focuses on comparing values of a single time series or multiple dependent time series at different points in time. The first model we used in this project is ARIMA, an acronym that stands for autoregressive integrated moving average.

2.3 Long Short Term Memory (LSTM)

In our project, LSTM is the second time series model. The LSTM means long short-term memory. A powerful type of neural network designed to handle sequence dependence is called recurrent neural networks (RNN for short) [3]. The LSTM model is a type of RNN, which can have the ability to handle non-linear time series prediction problems by feed-forward networks using fixed size time windows. Comparing to ARIMA, the promise of LSTM is that the temporal dependence in the input data can be learned. We used the LSTM model to learn the trend and the seasonal changes of a given data set more accurately.

3 Predication Results

The data tables given by Fastenal are ".csv" file format. The size of the 4 YEAR CONSUMPTION table is over 9 GB with over 146 million lines in the table. Because the computing limitation, we choose the smallest distribution center that has nine branches for this project. There are a total of 108, 535 items. Figure 1 is the screen shot of our web-based application. Please note, the consumption table is encrypted, hence the distribution center, branch and items are non meaningful strings in the figure.



Items in qpqxc

#	SRV_HUB	BRANCH_CODE	INV_ITEM_ID	LINEAR_REGRESSION	ARIMA	LSTM
26212	qp~<<	qpqxc	!!!{~	View details	View details	View details
26213	qp~<<	qpqxc	!!!~&	View details	View details	View details
26214	qp~<<	qpqxc	!!!~)	View details	View details	View details
26215	qp~<<	qpqxc	!!%~!	View details	View details	View details
26216	qp~<<	qpqxc	!!%~>	View details	View details	View details
26217	qp~<<	qpqxc	!!%~]	View details	View details	View details

Figure 1 The Screen Shot of the Web-based Sales Predication Application

The user can select any of the three machine learning algorithms to predict the future sales. Figure 2 is the screen shot of the Linear Regression prediction results.

LR_JUN2016	LR_JUL2016	LR_AUG2016	LR_SEP2016	LR_R2	LR_RMSE	LR_TIME
4.408	4.549	4.691	4.832	0.058	4.623	0.016
0.163	0.168	0.173	0.178	0.049	0.171	0.001
4.07	4.2	4.331	4.461	0.049	4.268	0.001
0.235	0.242	0.249	0.257	0.044	0.246	0.001
0.338	0.349	0.36	0.371	0.055	0.355	0.001
4.07	4.2	4.331	4.461	0.049	4.268	0.001
7.822	8.069	8.316	8.562	0.044	8.197	0.001
7.822	8.069	8.316	8.562	0.044	8.197	0.001
4.123	4.245	4.368	4.49	0.03	4.309	0.001
2.442	2.52	2.598	2.677	0.049	2.561	0.001
5.254	5.416	5.579	5.742	0.039	5.501	0.001
0.085	0.087	0.09	0.093	0.055	0.089	0.001
0.326	0.336	0.346	0.357	0.049	0.341	0.001
0.075	0.077	0.08	0.082	0.039	0.079	0.001
0.075	0.077	0.08	0.082	0.039	0.079	0.001
0.507	0.524	0.54	0.557	0.055	0.532	0.001
0.275	0.283	0.291	0.299	0.03	0.287	0.001
0.069	0.071	0.073	0.075	0.03	0.072	0.001
0.3	0.31	0.319	0.328	0.039	0.314	0.001

Figure 2 Linear Regression Sale's Predication Results

The 4-year consumption table has the sales data for 48 months. We chose data from 44 months as the training data, to predict the sales for the next four months. The actual data of the last four months was used to measure the accuracy of the prediction. To evaluate the forecast result, we are using the Coefficient of Determination (R^2 for short), and the Root Mean Squared Error (RMSE) to measure the accuracy of our model.

R^2 is a measure of the residual error in the model. We can use this as a measure of how good the model fits the historic data set. R^2 takes on values from (x, 1], with 1 meaning that the model is a perfect fit to the dataset, where x can be negative values. When fitting non-linear functions to the data, R^2 maybe negative.

RMSE is used to compare forecasting errors of different models for a particular data. The smaller of the value, the better the model. We used RMSE to measure which machine learning algorithm is better for a particular item.

Of the 108,535 items, approximately 30,000 items have R^2 values close to 1; the rest 70,000 items have R^2 values close to 0.1 because there are many zero sales and the data is unpredictable. We define the data is predictable in our project whenever R^2 greater than 0.8. However we still use Linear Regression because it is very efficient. This algorithm can process large amounts of data in a very short period of time, and without stressing computational ability. In Figure 2, the column LR_Time shows the actual execution time for the Linear Regression algorithm.

Now we chose a random item to compare three machine learning algorithms in detail. For item **--]]<<)]-~)~]>{**, Figure 3, 4 and 5 show the 4 months predication results, two accuracy measurement results and the run time of each machine learning algorithm.

```
Index:  11
BRANCH_CODE:  ay~<<
INV_ITEM_ID:  ]]<<)]-~)~]>{
>Predicted=1165.554, Actual=1080
>Predicted=1168.705, Actual=864
>Predicted=1171.856, Actual=972
>Predicted=1175.007, Actual=1296
RMSE: 196.689
R^2: 0.027
Runtime: 0.002 seconds
```

Figure 3 Linear Regression Result

```
>Predicted=1367.419, Actual=1080
>Predicted=881.347, Actual=864
>Predicted=1240.000, Actual=972
>Predicted=916.167, Actual=1296
RMSE: 273.408
R^2: -1.930
Runtime: 0.734 seconds
```

Figure 4 ARIMA Result

```
Predicted=1159.003, Actual=1080.000
Predicted=1250.427, Actual=864.000
Predicted=1034.727, Actual=972.000
Predicted=955.006, Actual=1296.000
RMSE: 262.573
R^2: -1.702
Runtime: 620.255 seconds
```

Figure 5 LSTM Result

For this particular item, linear regression has both the best R^2 and RMSE among the three models. But, for other items, ARIMA has the best R^2 value and LSTM has the best RMSE value. We are not able to draw conclusion for all the items, which model has the best accuracy because Fastenal data is encrypted, we lack of product demand and business knowledge. To our project, those data is unpredictable in nature.

In general, linear regression is the simplest predication model. However, this model is not be able to handle the complexity of the data. LSTM is currently the most promising model using a time series approach.

However, consistently in every item's running time, the linear regression is fastest, and the LSTM is the slowest. LSTM's significant runtime limits the application of this model due to the large amount of data using our existing computation power.

4 Proposed Predication Process

To apply the machine learning algorithms in practice to Fastenal consumption data, we propose the following predication process shown in Figure 6. We ran the data through a Linear Regression model first. Based on the model accuracy measurement, we either accept the forecast result or pass the bad fit data into the ARIMA model. Similarly, we then either accept the ARIMA model result or pass the bad fit data to the LSTM model.

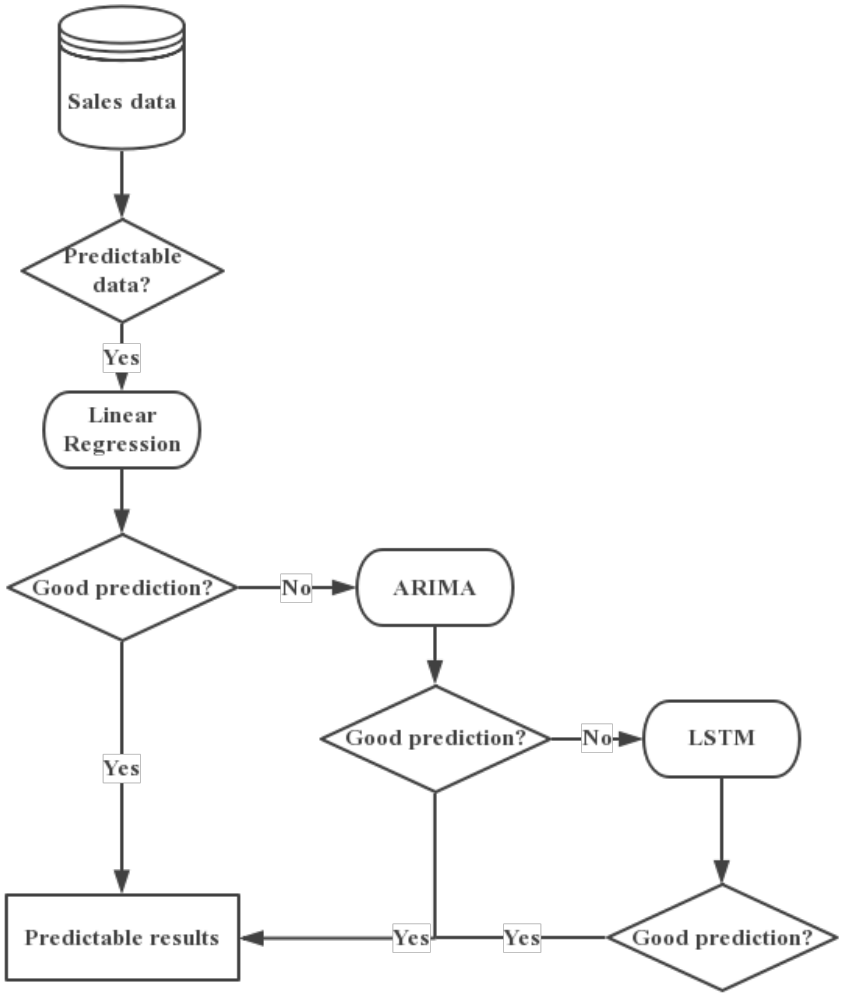


Figure 6 Predication Process

6 Conclusions

Linear regression is the simplest approach. It is efficient, but not be able to handle complex seasonal changes in product demand along with other factors. LSTM is currently a very popular model to explore complex time series. However it is significant slower than the other two models in our project. The next step of our research is to improve accuracy through data reorganization.

With forecasted sales, we can start to design the optimization of the Fastenal inventory. We will need to look at the distribution cost and storage cost and build a model to analysis the cost and benefits.

References

[1] https://en.wikipedia.org/wiki/Machine_learning.
[2] https://en.wikipedia.org/wiki/Time_series
[3] https://en.wikipedia.org/wiki/Recurrent_neural_network