2019

Optimizing SIG's sales and operations planning by forecasting customer demand for packages in Thailand market

Class: NTHU, Business Analytics using Forecasting

Instructor: Galit Shmueli

Group Members: Pei Pei Chen, Lynn Pann, Amber Lin, Saskia Thus

Jan 2019

Executive summary

This project is performed on behalf of SIG. SIG is a global company that produces packages and filling machines for food and beverages. SIG has multiple different clients within Europe and Asia. For each of these customers, SIG customizes the package in terms of size and print. Because of this customization, SIG has a lot of different “products” (combination of a customer and one of his requested package materials) to fulfill. For the sales and operation planning purpose, SIG’s salespersons need to hand in the forecast of demand for different products every month. However, the forecasts made by the salesperson are not accurate enough and therefore the planning has to be adjusted during the month, which costs money. Therefore, this project is helping SIG forecast the sales of these customized customer package materials. In this report, the focus is placed on the customers of SIG in Thailand. However, the final delivered model can also be used for customers from other countries.

The data set came from SIG, After data processing and filtering by country Thailand, SIG has 20 customers in Thailand and 85 different “products” (combination of a customer and one of his requested package materials) of time series in total. Before the actual forecasting took place, the data is analyzed and the customer packages are categorized into three categories: active customers, past customers, and new customers. New customers are customers started to have demand for their package in the last 2 years, past customers are customers that stopped having demand for their package in the last 2 years and active customers are customer packages that already have demand for more than 2 years and had demanded in the last 2 years as well. Only for the active customers' forecasts are given for the upcoming 12 months. For the remaining two categories, either they don’t have any demand anymore so forecasts are not necessary or there is not enough data available to make accurate forecasts.

To forecast the sales of the customized customer packages for the active customers, two different models are built. A quick model is built using moving average with a window of 12 periods. Next, the best model is built for every customer package pair. This best model is the best out of five different forecasting models and is chosen based on the lowest RMSE and the forecast error distribution. The five used models are; seasonal naive, moving average, Arima, exponential smoothing and linear regression. Both the quick and the best model can give a forecast for the upcoming 12 months, where the best model can give more accurate forecasts and the quick model can give forecasts in less time.

Lastly, the recommendation is given to SIG about how they can improve the accuracy of the forecasts of every customer package. We believe this accuracy can be increased by using the forecasts that are currently given by someone of the sales department for the upcoming 12 months as external data into the automated models we build.

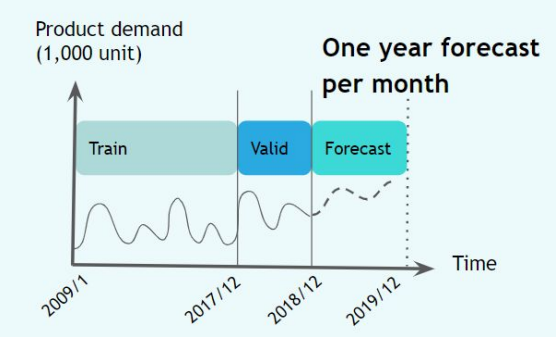
Problem Description

Business goal

The stakeholder in this project is the operations department of SIG. They are responsible for creating the sales and operations planning and therefore take the biggest advantage of the more accurate forecasts. Currently, SIG already makes use of forecasts for the sales volume of every customer package type on a monthly basis. These forecasts are made by salespersons and are handed before the 18th of every month. Based on these forecasts the operations department of SIG makes a sales and operations planning for the upcoming month(s). However, it turns out that the forecasts made by the salesperson are not accurate enough and therefore the planning has to be adjusted during the month, which costs SIG money every month. Therefore, the business goal is to optimize the demand forecast for sales and operation planning monthly every year for operation department. When the sales and operation planning of SIG is optimized, the entire company will benefit for two reasons. First of all, by optimizing the operations planning more demand can be fulfilled. Next, by efficiently producing the financial pressure will decrease for SIG.

Forecasting goal

The forecasting goal of this project will be to forecast the number of customer packages for every customer package separately for the upcoming 12 months in thousands of units. This will be done once a year on December 18th for the 12 months in the upcoming year. This forecasting goal is a forward-looking goal because the information in the received time series from SIG will be used to forecast future values of that series.



Data Description

Data

The raw dataset is customers' orders from years 2009 to 2018 provided by SIG with columns of Customer, Country sold-to party, Company code, Material, Material Description, Product hierarchy, Plan quantity, and Month. After data processing and filtering by country Thailand, SIG has 20 customers in Thailand and 85 different "products" (combination of a customer and one of his requested package materials) of time series in total. For every package of a certain size and with a certain print, SIG records the total number of units (in thousands of units) of that package that was sold in each month. Each series is in different length. So for example when a package was first ordered in January 2016, data is given for each month from January 2016 till December 2018, so 36 data points. These data points together will be called a time series from now on.

Products (cust + package)	200901	200902	200903	200904	200905	200906	200907	200908	200909	200910	200911	200912	201001
400005 PC031000J	161.4	0	163.8	0	0	0	0	0	0	0	0	0	0
400008 PC010125A	0	0	0	0	0	0	0	0	0	0	0	0	0
400008 PC010125J	0	0	0	0	0	0	0	0	0	0	0	0	0
400008 PC010200J	0	0	0	0	0	0	0	0	0	0	0	0	0
400008 PC010250J	0	0	0	0	0	0	0	0	0	0	0	0	0
400008 PC020500J	0	0	0	0	0	0	0	0	0	0	0	0	0
400008 PC020750J	0	0	0	0	0	0	0	0	0	0	0	0	0
400008 PC021000J	2221.8	2291.1	799.2	193.8	1787.7	1501.5	2868.6	4209	1323	965.4	2849.7	448.2	2256.9
400008 PT021000J	0	0	0	0	0	0	0	0	0	0	0	0	0
400013 PC010125A	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 1. Screenshot of time series dataset

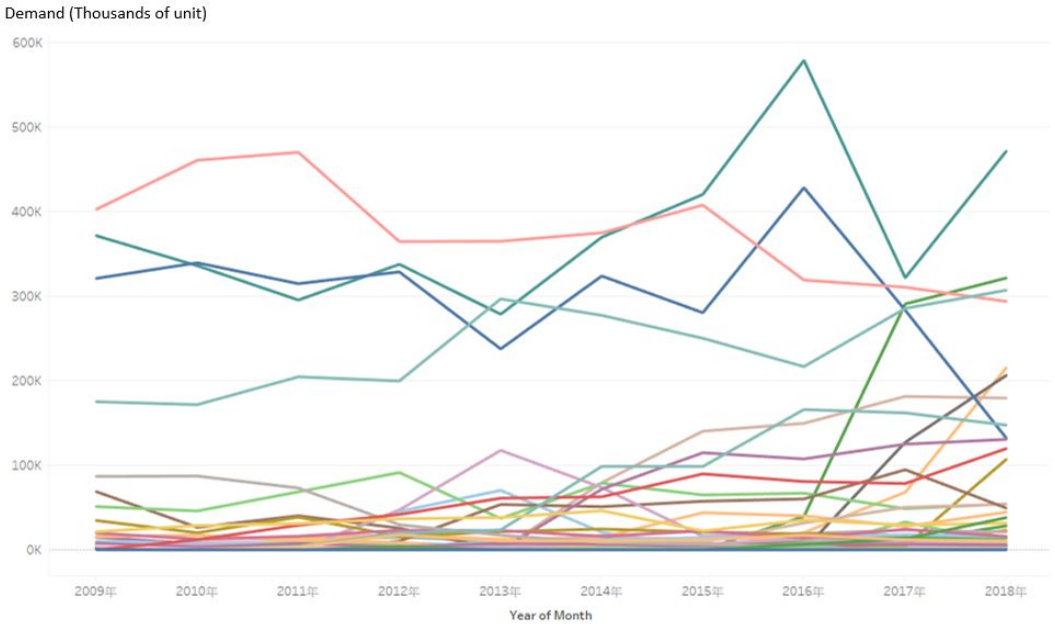


Figure 2. Graph of demand for every package between 2009 and 2018

Data exploration

Before we started working with the model, we explored all the time series. To do so, we tried to analyze how hard it would be to forecast the sales of each of the packages. For this, we run the automated exponential smoothing function in R over the data. This function returns for every of the time series whether or not this time series shows a trend, seasonality or both. A trend is the tendency of the data to either gradually increase or gradually decrease over time. Seasonality means that during a season the data shows the same pattern. This can, for example, be in this data that the sales of the packages are every year higher in December than in the rest of the year.

The analysis of the data showed us that 61% of all packages showed no seasonality or trend, 34% only showed a trend and 5% only showed seasonality. The forecasting models known in the literature are mostly more easy to apply to data with either a trend or seasonality. Therefore, this analysis has shown us that 61% of the data will be hard to forecast and that we should be prepared for the fact that low accurate forecasts may be given for these packages.

Brief data preparation details

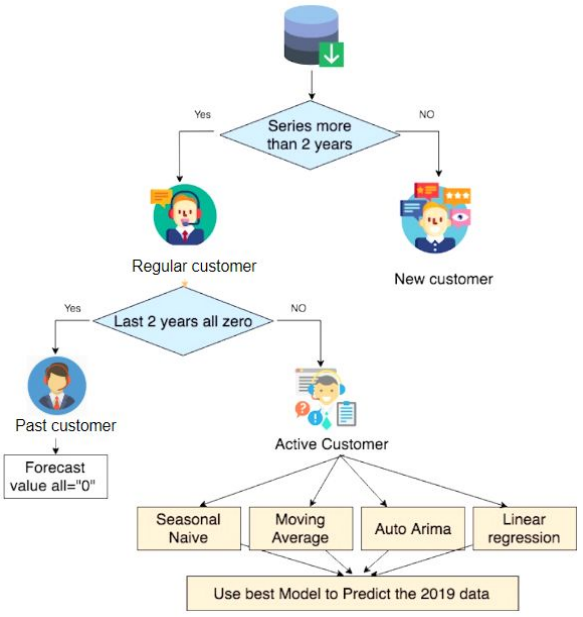
Data processing

From SIG's manager's suggestion, Plan quantity that is negative can be ignored from the raw data. So we change all the negative value into zero before we aggregating. Next, we aggregate all the Plan quantity from the raw data into the same customer and month. Last, we turn all the zero value before the first-month order into Null value in each time series. These give us the different length of time series.

Data segmentation

After data processing, we segmented the time series into three types; packages of new customers, past customers, and active customers. The first step in this flow is to determine if for the time series of that package contains more than two years of data. So if that package is already sold for more than two years or not. If it is not, the packages will be classified as a new customer and no forecasts will be made for this package. This decision is made in order to make an accurate forecast for a package, so enough old data should be available to base the forecast on. If not enough data is available, the forecast will be just a "random" guess.

When more than two years of data is available for that package, a new check is executed. This check is whether or not the sales of this package were all zero over the last two years. If this is the case, the package will be classified as a past customer and the forecast for the next 12 months will be zero for each month as well. If there was some sale over the last two years, the packages are classified as an active customer. In that case, four different models are tried to fit on the time series, the different methods will be explained in the next section in detail. The best model of these four models will be used to predict the sales for the next year, which is 2019 when this model was built.
Based on the given data from 2009 till 2018, 13 packages were classified as being new customers, 38 packages as being past customers and for the remaining 34 packages, one of the four models was fit to the data.



Forecasting solution

Model building

To forecast the demand for the upcoming 12 months, we first split the data into a training period and a validation period. Based on the training period the method will fit a model and this model will then be tested in the validation period. The results of the forecast for the validation period will be compared with the actual demand in the validation period. The data from 2009 to 2017 is used as the training period, where the data of 2018 is used as the validation data.
To forecast the sales in 2019, we build two different models. The first model is a model which compares the results of five different methods with each other and chooses the best of these methods to actually forecast the sales in 2019. The best model is chosen based on the RMSE. The five methods that are compared to choose the best model are the moving average, exponential smoothing, linear regression, auto Arima and seasonal naïve forecast. For the moving average, a window of 12 months is used, for the seasonal naïve a season of 12 months is used and for all other methods, the automated function in R is used. This automated function determines the value for each parameter itself. The second model is called the quick forecast and only makes use of one method; the moving average. By only making use of only one method, this model can give a forecast faster than the best model.
Both models are compared to our benchmark model, which is the seasonal naïve forecasting model. We expect both models to beat this benchmark model. This last comparison between the benchmark model, the best model, and the quick model will be done based on the forecast error plot. SIG indicated that they prefer to have under forecasts than over forecasts and therefore we will look in the forecast error plots for positive forecast errors.

Performance evaluation

As told before, the different models are evaluated based on their RSME and on their forecast error plots. For each of the models, the performance is shown in Figure 4. When we compare the best model with the benchmark model, we can safely say that the best model beats the benchmark model. The RSME is about 20% lower and also the forecast error plot shows us that most of the forecast errors of the best model are positive, where most forecast errors of the benchmark model are negative. Since SIG wanted us to underpredict, this shift of the errors from negative to positive is an improvement.
Next, we compare the quick model with the benchmark and best model. The RMSE is increased a bit in comparison with both other models, but the distribution of the forecast error is again positive. The fact that the

RMSE is higher for the quick model than it is for the best model is a trade-off with the running time of the model. Where the running time for the best model is around 1.5 to 2 minutes, the running time for the quick model is only half a minute.

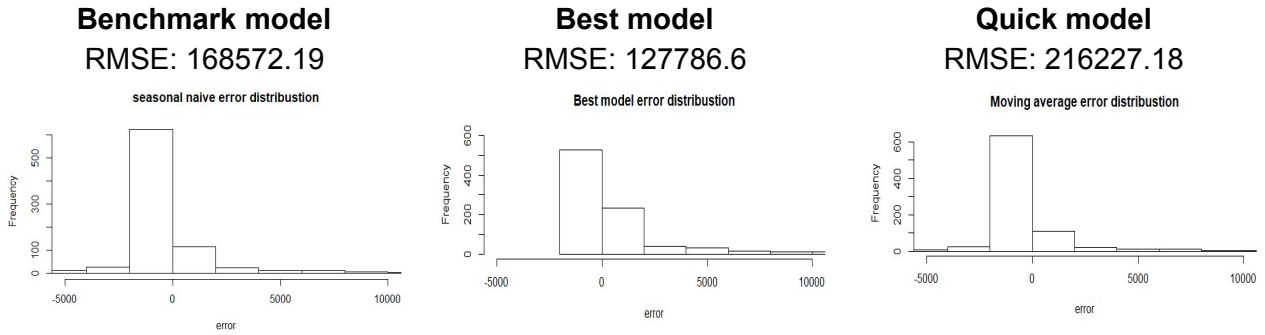
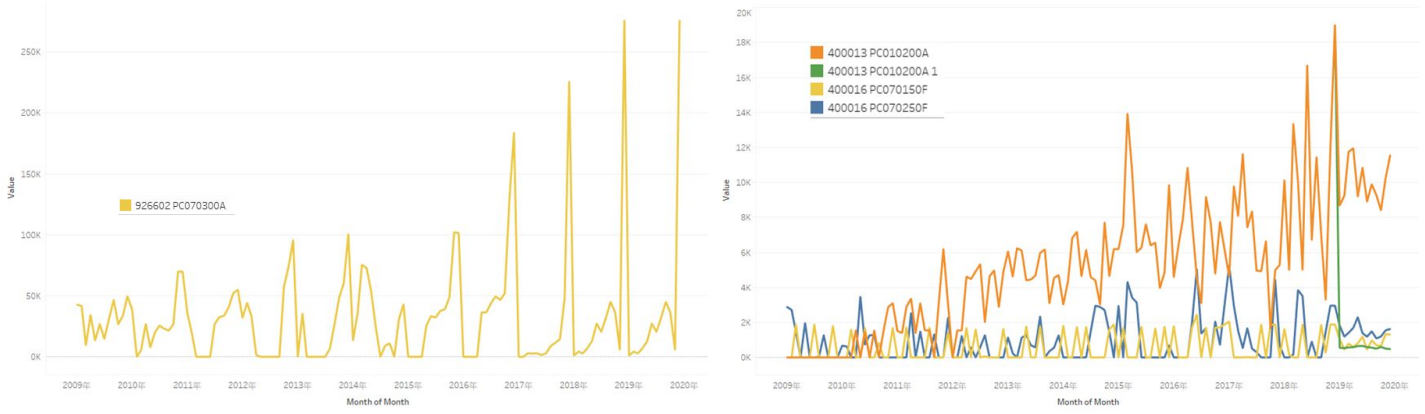


Figure 4. Performance and results

Unfortunately, we were not able to compare the forecasts of the best model with the forecast provided by the sales manager. This can be something that should be compared in the future.

Time plot of series with future forecasts (2019)



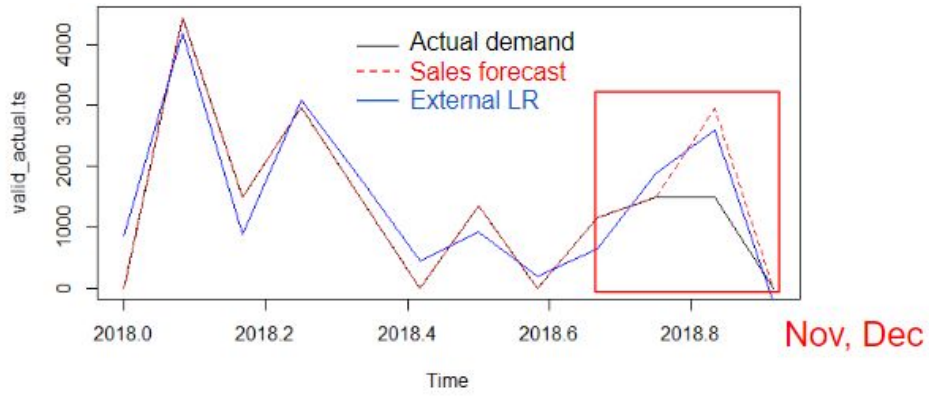
Conclusions

As explained before, the goal of this project was to find better forecasts for the sales volumes of the different packages of the different customers of SIG in Thailand, than the salesperson provides nowadays. So far in this report, we discussed different forecasting models that are based only on the actual sales and that (unfortunately) couldn't give accurate results for every package. A suggestion that we want to give to SIG is to compare the forecasting models we build with the forecasts provided by the salesperson. This can easily be done in R as well by using the function for external linear regression. This function still does linear regression in the way explained before, but now the formula is a bit adjusted and also external data is taken into account. (see more in Appendix: Future work). We also suggest using the roll-forward validation method for giving more evaluation data.

However, there are some limitations to our models. First, this is an ongoing analysis requiring collecting new data. Second, we heard from SIG that their policy changed every two years including discount and product generation. This influences the orders and the forecasts. So, the advantage of our model is that it will automatically select the best model for forecasting next year. Third, the limit amount of sales data is close to the forecasting goal (k=12) months, the more data is added to the time series, the forecast performance would become more stable.

APPENDIX: Future work

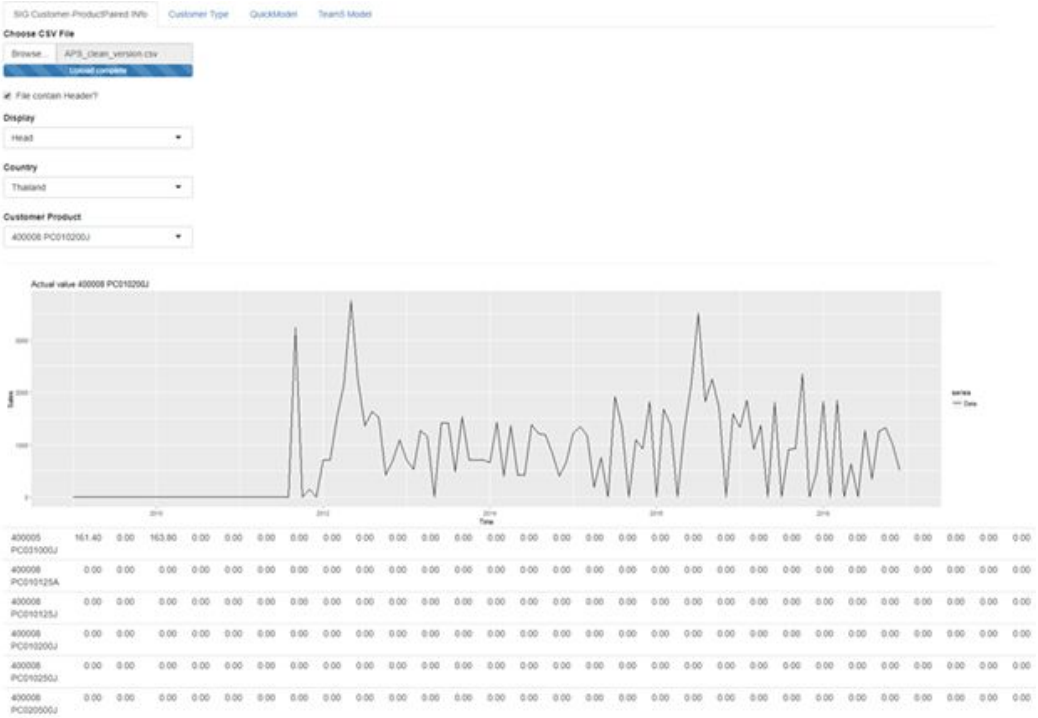
The graph below shows in which the forecasts of the salesperson are taken as external input for the forecasting model we build. The black line is the actual demand, the red line the sales forecast provided by the salesperson and the blue line the forecast of our forecasting model which uses the actual sales over the last months and the sales forecast of the salesperson into account. As can be seen, the forecasting model (blue line) predicts the sales better than the salesperson (red line). Unfortunately, we were not able to test this on more months than only on November and December of 2018, but we believe that the combination of the two models will increase the accuracy of the forecasts in other months as well.



APPENDIX: Interface

In this last section, we will give a short manual about how to use our model for forecasting the sales volume of the different packages for every customer. To make the model easier in use, we build an interface, in which you only have to select what you want the model to do and it will do it for you. No additional knowledge is needed from coding or forecasting to use the interface. Next, a short manual will be given about where to click in the interface to find the forecasts.

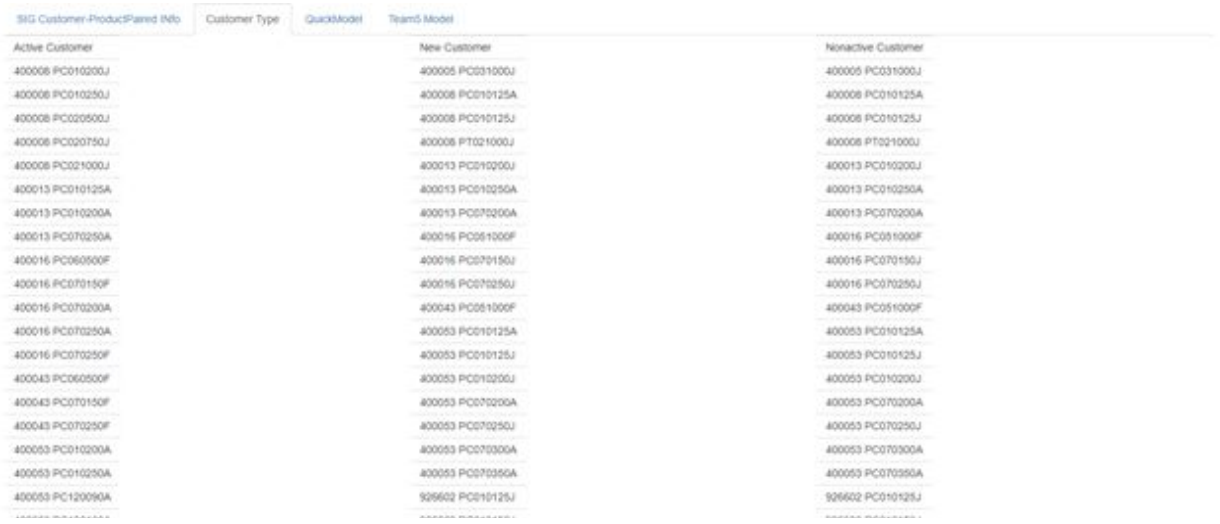
SIG Customer-ProductPaired Info page



Step 1: select the data file with the actual sales of every customer's package.
Step 2: There are three things you can select now; display, country and customer product.

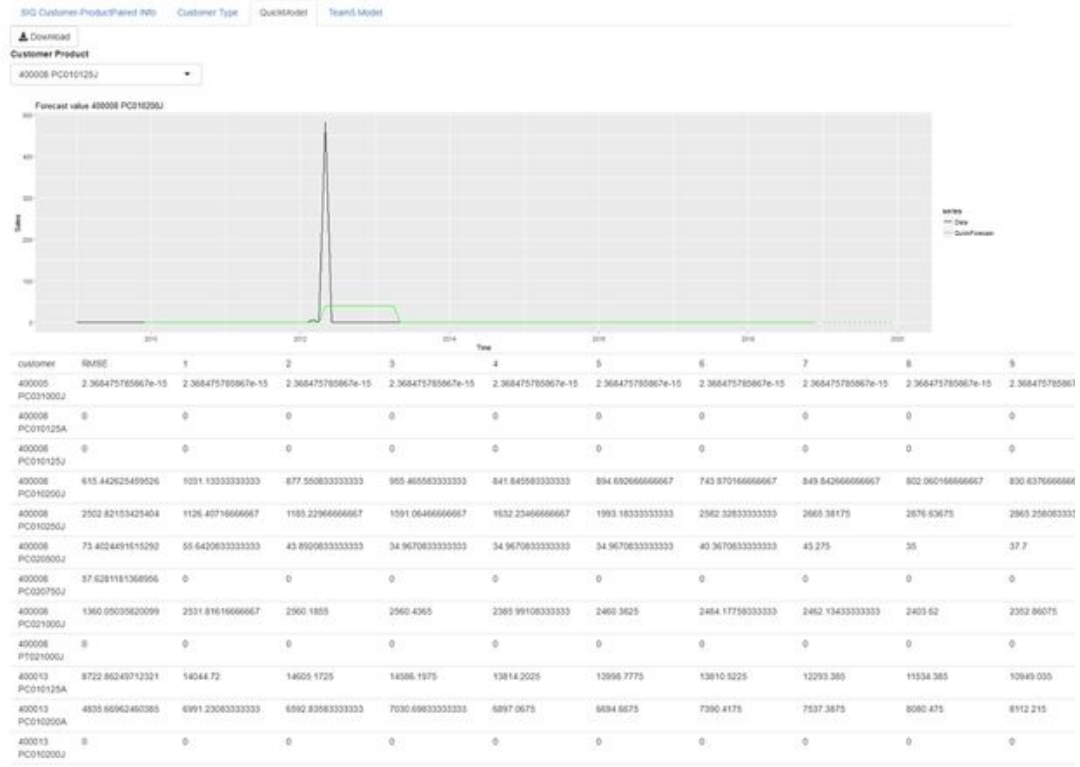
- The display option should be left to Head.
- In the Country option, you can choose the country for which you want to forecast the sales volume in the upcoming 12 months.
- In the Customer Product option, you can choose which customer package pair you want to see in the graph below. That graph will show the past data from the sales of that customer package pair that is provided in the selected excel file. Under the graph, you find a table with the sales of every customer package pair that is provided in the selected excel file...

Customer Type page



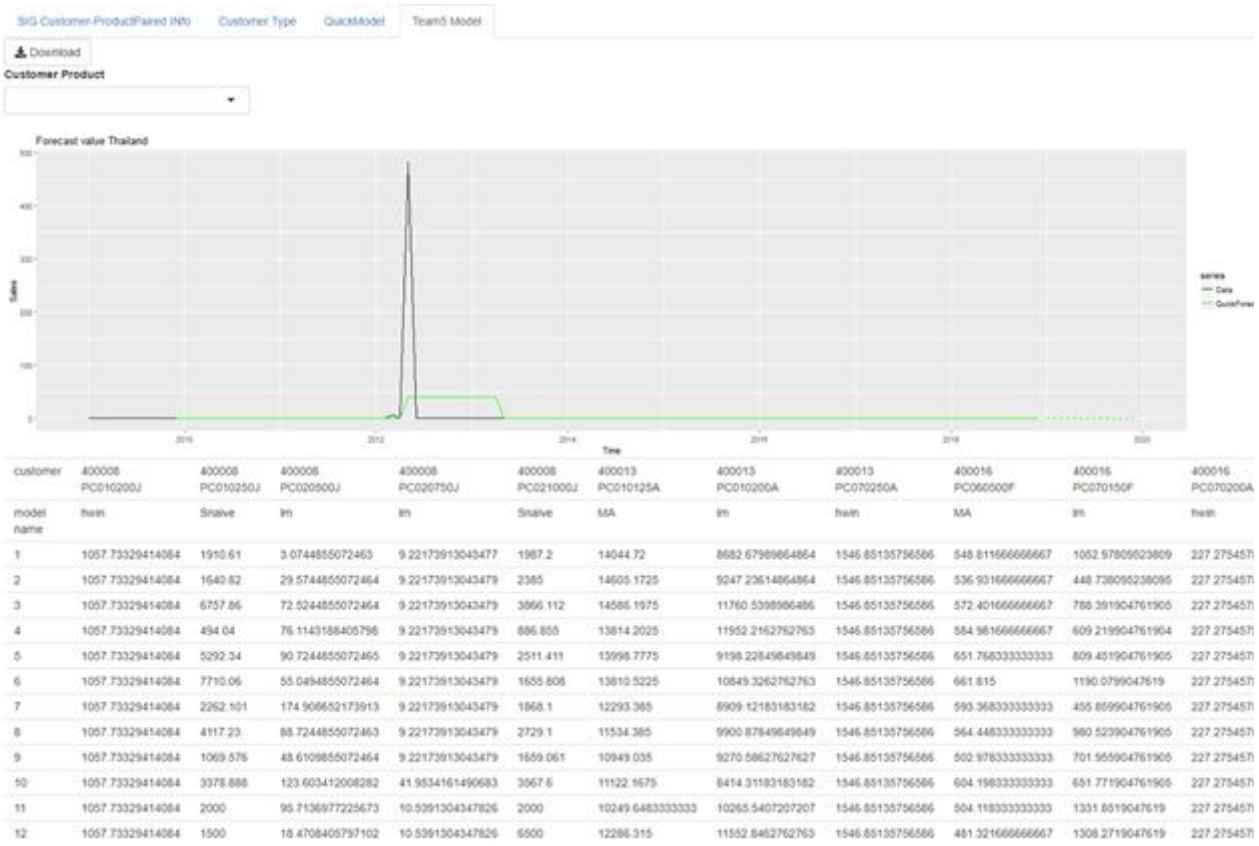
On this page, you find which customer packages are selected to be active customers, new customers or nonactive customers by the model. For all the new customers, no forecasts are given, for all the nonactive customers' forecasts of all zeros are given for the upcoming 12 months and for the active customers a model is built, which forecasts the sales in the upcoming 12 months.

Quick model page



On this page, the quick model is presented, which is the model that bases his forecasts of the sales volume on the moving average with a window of 12 months. Again, you can choose a customer product, which will be displayed in a graph below the drop option. In the graph, the black line is the actual sales in past periods and the green line represents the forecasted sales. Below the graph, you can find a table with the forecasts for the sales of each customer package in the upcoming 12 months. The first column displays the customer package, the second row the RMSE for the forecasts of that customer package and the columns 3 till 14 represent the forecasts in the upcoming months. So, for now, the input data was 2009 until 2018, so the forecasted sales are for 2019. Therefore the third column with a 1 in the top represents January 2019 and so forth.

Team5 model page



customer	400008 PC010200J	400008 PC010250J	400008 PC020500J	400008 PC020750J	400008 PC021000J	400013 PC010125A	400013 PC010200A	400013 PC070250A	400016 PC060500F	400016 PC070150F	400016 PC070200A
model name	hwin	Snaive	lm	lm	Snaive	MA	lm	hwin	MA	lm	hwin
1	1057.73329414084	1910.61	3.0744855072463	9.22173913043477	1987.2	14044.72	8682.67989864864	1546.85135756586	548.811666666667	1052.97809523809	227.275457
2	1057.73329414084	1640.82	29.5744855072464	9.22173913043479	2385	14605.1725	9247.23614864864	1546.85135756586	536.931666666667	448.738095238095	227.275457
3	1057.73329414084	6757.86	72.5244855072464	9.22173913043479	3866.112	14586.1975	11760.5398986486	1546.85135756586	572.401666666667	788.391904761905	227.275457
4	1057.73329414084	494.04	76.1143188405796	9.22173913043479	886.855	13814.2025	11952.2162762763	1546.85135756586	584.981666666667	609.219904761904	227.275457
5	1057.73329414084	5292.34	90.7244855072465	9.22173913043479	2511.411	13998.7775	9198.22849849849	1546.85135756586	651.768333333333	809.451904761905	227.275457
6	1057.73329414084	7710.06	55.0484855072464	9.22173913043479	1655.808	13810.5225	10849.3262762763	1546.85135756586	661.815	1190.0799047619	227.275457
7	1057.73329414084	2262.101	174.906652173913	9.22173913043479	1868.1	12293.385	8909.12183183182	1546.85135756586	593.368333333333	455.859904761905	227.275457
8	1057.73329414084	4117.23	88.7244855072463	9.22173913043479	2729.1	11534.385	9900.87849849849	1546.85135756586	564.448333333333	980.523904761905	227.275457
9	1057.73329414084	1069.576	48.6109855072464	9.22173913043479	1659.061	10949.035	9270.58627627627	1546.85135756586	502.978333333333	701.959904761905	227.275457
10	1057.73329414084	3378.888	123.603412008282	41.9534161490683	3567.6	11122.1675	8414.31183183182	1546.85135756586	604.198333333333	651.771904761905	227.275457
11	1057.73329414084	2000	95.7136977225673	10.5391304347826	2000	10249.6483333333	10265.5407207207	1546.85135756586	504.118333333333	1331.8519047619	227.275457
12	1057.73329414084	1500	18.4708405797102	10.5391304347826	6500	12286.315	11552.8462762763	1546.85135756586	481.321666666667	1308.2719047619	227.275457

This page presents the best model fitted among the moving average method, the exponential smoothing method, the linear regression method, the Arima method, and the seasonal naïve method. This page works the same as the quick model page, where you can choose your customer package that you want to be displayed in the graph and the black line represents the actual sales and the green line the forecasts. And again below the graph, the table is presented with the RMSE and forecasts for the upcoming 12 months for every customer packages