BAFT 2018

Forecasting swimsuit sales for the next month to assist in inventory management for Heatwave

Team 4

Jheng Kai-Ru

Adam Yu

Silvia Yang

Zoly Chang

National Tsing Hua University, Institute of Service Science

Executive Summary

Heatwave is a swimsuit e-commerce seller on the Tmall.com. They designed and manufactured their swimsuits and sell them on the Tmall platform. In this project, our team collaborated with Heatwave to work on the forecasting job of predicting swimsuit sales for the next month to assist in inventory management.

The goal of Heatwave is to estimate the future inventory and decide the number of products to put into manufacture for the next month. Moreover, they would also like to adjust their marketing strategies through forecasts. For instance, they would put promotions on the products if the forecasts are lower than their expectation.

In the beginning of the project, the data pre-processing took us lots of time and we made a big effort on integrating two sources of data. The data quality was low and inaccurate. Eventually, we got only 19 months of sales data in total.

We then focused on the top 5 products and generated forecasts using both monthly and daily data. The seasonal naive model was chosen as a benchmark to compare with other models. We found that seasonal naive actually worked best for the monthly data. On the other hand, the arima and ets model performed well in the daily data.

In the implementation phase, we ran into several difficulties. First, Heatwave lost the majority of their historical data during the platform transition, causing the short length of data. Secondly, the low quality and inconsistency of the data made it hard to pre-process and integrate, which cause some losses of data. Thirdly, we were unable to forecast newly released products due to the short data length. Lastly, the characteristic of the short product cycle in the swimsuit industry makes the dataset relatively short in nature, the forecast will work only if the data can be processed and delivered immediately.

In summary, we suggest Heatwave do the followings:

(1) Save the data periodically to retain the data autonomy (ownership) and data quality.
(2) We found similar patterns between new and old products. The company can use the sales of old products as part of the ingredient when forecasting new products with a similar pattern.
(3) Modify models as the datasets become longer. The company can have more precise insights into the product life cycle. Eventually, reaching the goal of lean production.
(4) Make decisions not only with the forecast value but also manager’s domain knowledge.

Problem Description

Background

The company we collaborated with is Heatwave, a B2C swimsuits seller on a Chinese e-commerce platform called TMall. They self-designed and self-manufactured their own swimsuits. Our forecast job is to generate forecasts of the swimsuit sales for the next month. The biggest challenge Heatwave is currently facing is that they had limited knowledge about how much to manufacture for the next time and when to put promotions on certain products. Their decisions were mostly made upon their past experiences and domain knowledge. However, our client experienced a data loss around one and a half years ago when switching platform, the data we received is relatively short, which became our biggest obstacle in this project.

Business Goal

As a result, the business goal of Heatwave is two-fold:

1. Adjusting the amount of manufacture, and managing inventory in advance
2. Deciding whether to put promotions on certain products according to the forecasted sales each month

If the forecasted sales of the products next month are lower than the current month, they might adjust the amount of manufacture or put promotions on it on the TMall platform. If a product is already on promotion but the forecasted sales are still low, they might consider putting more promotion on it or simply take off all the promotions and focus on other product.

The forecasting result will be considered successful if we accurately forecast the sales numbers for the next month. Besides, Heatwave prefers a slight under-prediction than over-prediction, which may trigger promotional events more easily and keep the inventory low, which aligns with the goal of lean production.

Forecasting Goal

Our forecasting goal is to forecast the total sales of the certain type of swimming suit for the next month, which is a predictive goal. The time index “t” represents the month and “t=1” represents the first-month data are able to acquire in the series. “yt” represents the actual sales number at the time “t”. The forecast horizon “k” is 1 month. The forecast result is expected to be updated at the end of every month. The data is updated daily, so at the time of prediction, we can get last month sales numbers.

Data Description & Pre-processing

Our data came from two sources: Sheng-Yi-Can-Mou and TMall's analytic platform. By integrating these two sources, we acquired sales data from April 2017 to November 2018 at both daily and monthly scale. During the integration, we ran into several difficulties, including the inconsistent data format from two sources and product identification codes that do not match. The detailed data problems are listed in the **appendix**. We present the problem of data quality and the integration of data down below.

Data Quality And Limitation

We hoped to forecast the sales of the next month by integrating the two sources of data. However, the data from TMall's analytic platform is not very reliable and precise. The low data quality brought many difficulties when combining two datasets. The followings are the description of the difficulties.

To begin with, we could only acquire monthly data after Nov 2017 and daily data after Aug 2018 in the Sheng-Yi-Can-Mou while we have the data from April 2017 to November 2018 in the TMall's analytic platform. Then, we found that the number of orders and products are inconsistent in two sources, hence we could not directly combine two datasets. Furthermore, product names are displayed differently in two sources and some products are with different product identification numbers. We found 1314 records that are missing product identification number; hence are regarded as missing values.

As a matter of the above reason, our data quality from April 2017 to July 2018 is relatively low due to the shortage of data. However, we still need these data to help us generate forecasts. Despite some inaccuracy of the data, we still find some sales patterns in the data and we think it's very helpful.

Integrated Data

Result from many missing and incomplete data occurred during the integration, we chose five products that have a relatively more complete dataset. We look at the data in both daily and monthly scale to capture different patterns of data. The daily data starts from 2017-04-22 to 2018-12-18 with a period of 7 days and monthly data starts from 2017-05 to 2018-11 with a period of 12 months.

Monthly Data

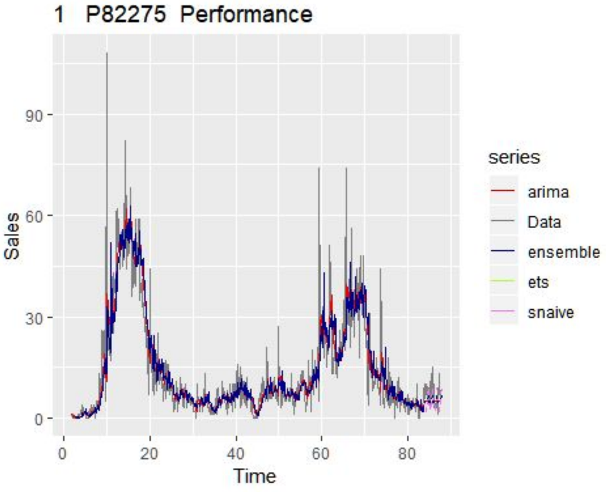
Daily Data

Forecast Solutions

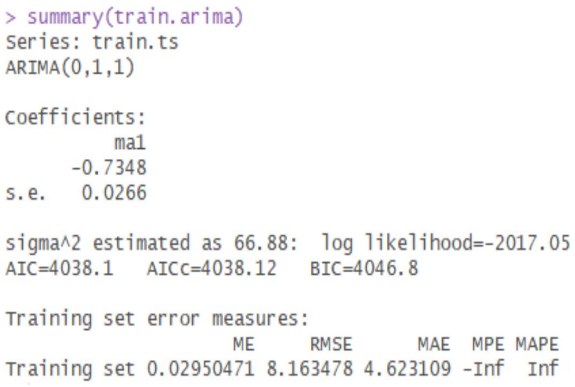
We used 48 days for daily data (from 2018-11-01 to 2018-12-18) as the validation set and use the rest of the data for the training set. During the exploration phase, many forecasting models were considered and applied on the dataset and compare forecast errors of each model (all models will show in evaluation table). We try the dataset at both daily scale and monthly scale to capture both **weekday-weekend** and **winter-summer** patterns. We found out that the model in monthly data is overfitting and is easily to be over-forecasted. Considering the business goal, Heatwave would rather under-forecast the sales rather than over-forecast. Therefore, we thought the daily forecast will more suitable for them to conduct. The daily forecast is shown down below. For the monthly forecast, please refer to the attached **appendix**.

We discovered the sales on Nov 11 and Dec 12 are extremely high compared to the neighboring days due to the platform's "double eleven" and "double twelve" campaigns, thus we removed these outliers from the daily data. We used seasonal naive as the benchmark.

Daily Forecast Result Plot



The best model: ARIMA



```
> summary(train.arima)
Series: train.ts
ARIMA(0,1,1)

Coefficients:
          ma1
      -0.7348
s.e.   0.0266

sigma^2 estimated as 66.88:  log likelihood=-2017.05
AIC=4038.1   AICc=4038.12   BIC=4046.8

Training set error measures:
                     ME     RMSE      MAE  MPE MAPE
Training set 0.02950471 8.163478 4.623109 -Inf  Inf
```

Evaluation

Daily RMSE	Naive	sNaive (Benchmark)	ets	regression	arima	nnetar with external data	ensemble (snaive, ets, arima)
training	6.28	4.54	**3.32**	5.54	**3.29**	10.72	3.72
test	28.8	1.51	**1.00**	3.64	**0.98**	16.49	1.16

According to the forecast chart and RMSE table, we can infer that for the daily data, the ets model and arima model perform equally, because the ETS(A, N, N) and the ARIMA(0, 1, 1) are very similar to each other. We recognized some high peaks in daily data are not captured by the models. The ensemble doesn't work because all models tend to under forecast.

Limitations

Our biggest limitation is the shortage of data. It's pretty difficult to forecast monthly sales with only 19 months of data, especially when there's a clear summer-winter seasonal pattern. Moreover, the swimsuit product has a short product cycle. It's natural that swimsuit products usually won't sell more than 2 or 3 years except for classic types; hence resulting in short periods of data. On the one hand, products with a longer period of data might be out of the market soon. On the other hand, new products usually do not have enough data to generate accurate forecasts. Furthermore, we are not able to forecast fashion trends.

Recommendations

In summary, we have the following recommendation for our client - Heatwave:

1. Use the daily data with a forecast horizon of 30 days and aggregate them to form a forecast of total sales for the next month. If Heatwave could renew their data on the daily bases, then they might have a in-time database with high data quality. Which would help with their forecast as well as to lower the horizon and make a more precise forecast.
2. Save the sales data at the local server periodically to retain the data autonomy (ownership) and quality.
3. Compare similarities between newly released product and the old product. There might be similar patterns that are helpful for decision making.
4. If the data is long enough in the future, Heatwave can probably have the better understanding of the product life cycle, which can help with the inventory management, with the goal of reaching lean production.
5. Incorporate the domain knowledge into the forecast results in order to make better decisions.

Appendix

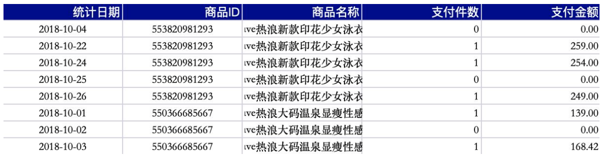
Integrating Two Sources of Data

Our final data for forecast are integrated from Sheng-Yi-Can-Mou and TMall's analytic platform. During the integration, we ran into several difficulties, including the inconsistent data format from two sources and data that don't match. Here we present the detailed data descriptions in two sources.

Data from Shen-Yi-Can-Mou

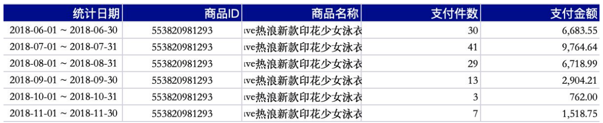
We collected data from the top 16 products in recent years, including daily sales, monthly sales, product names, and product IDs. However, we only have daily data from 2018-07-28 to 2018-12-18, which is a very short series to generate a useful forecast.

Daily data from Shen-Yi-Can-Mouê



统计日期	商品ID	商品名称	支付件数	支付金额
2018-10-04	553820981293	ıve热浪新款印花少女泳衣	0	0.00
2018-10-22	553820981293	ıve热浪新款印花少女泳衣	1	259.00
2018-10-24	553820981293	ıve热浪新款印花少女泳衣	1	254.00
2018-10-25	553820981293	ıve热浪新款印花少女泳衣	0	0.00
2018-10-26	553820981293	ıve热浪新款印花少女泳衣	1	249.00
2018-10-01	550366685667	ıve热浪大码温泉显瘦性感	1	139.00
2018-10-02	550366685667	ıve热浪大码温泉显瘦性感	0	0.00
2018-10-03	550366685667	ıve热浪大码温泉显瘦性感	1	168.42

Monthly data from Shen-Yi-Can-Mouêê



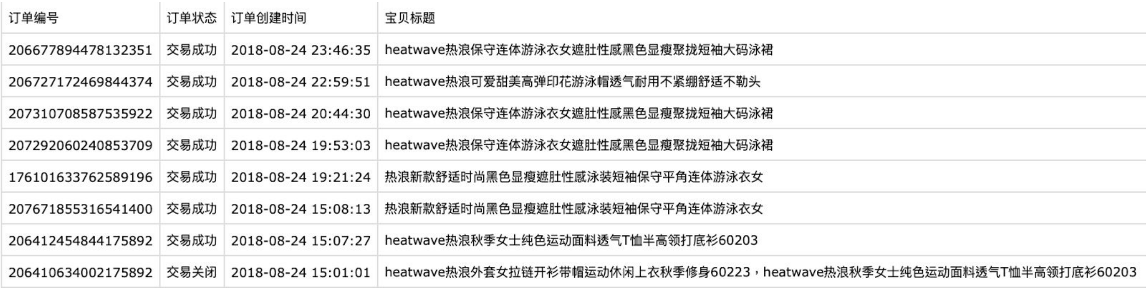
统计日期	商品ID	商品名称	支付件数	支付金额
2018-06-01 ~ 2018-06-30	553820981293	ıve热浪新款印花少女泳衣	30	6,683.55
2018-07-01 ~ 2018-07-31	553820981293	ıve热浪新款印花少女泳衣	41	9,764.64
2018-08-01 ~ 2018-08-31	553820981293	ıve热浪新款印花少女泳衣	29	6,718.99
2018-09-01 ~ 2018-09-30	553820981293	ıve热浪新款印花少女泳衣	13	2,904.21
2018-10-01 ~ 2018-10-31	553820981293	ıve热浪新款印花少女泳衣	3	762.00
2018-11-01 ~ 2018-11-30	553820981293	ıve热浪新款印花少女泳衣	7	1,518.75

Data from TMall's analytic platform

After discussing with Heatwave, we found more data from another platform. However, because they had a platform transition, we couldn't get data before May 2017. We acquired two lists of data that are related to product sales.

Order list dataê

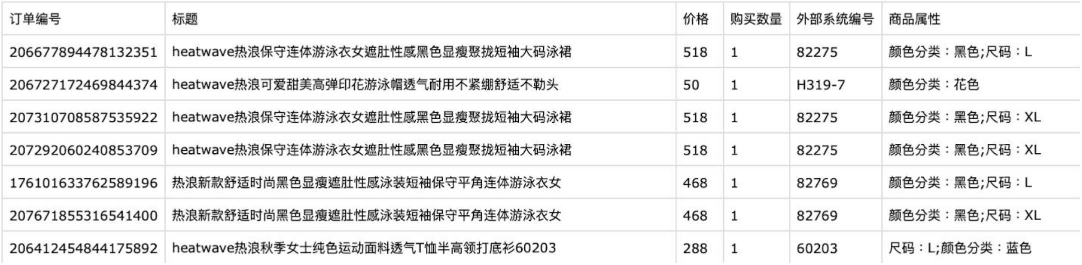
The order list includes the order number, timestamps for each order, products names in an order, price, and the number of products in an order. There might be multiple product names in an order and lack of product identification number.



订单编号	订单状态	订单创建时间	宝贝标题
206677894478132351	交易成功	2018-08-24 23:46:35	heatwave热浪保守连体游泳衣女遮肚性感黑色显瘦聚拢短袖大码泳裙
206727172469844374	交易成功	2018-08-24 22:59:51	heatwave热浪可爱甜美高弹印花游泳帽透气耐用不紧绷舒适不勒头
207310708587535922	交易成功	2018-08-24 20:44:30	heatwave热浪保守连体游泳衣女遮肚性感黑色显瘦聚拢短袖大码泳裙
207292060240853709	交易成功	2018-08-24 19:53:03	heatwave热浪保守连体游泳衣女遮肚性感黑色显瘦聚拢短袖大码泳裙
176101633762589196	交易成功	2018-08-24 19:21:24	热浪新款舒适时尚黑色显瘦遮肚性感泳装短袖保守平角连体游泳衣女
207671855316541400	交易成功	2018-08-24 15:08:13	热浪新款舒适时尚黑色显瘦遮肚性感泳装短袖保守平角连体游泳衣女
206412454844175892	交易成功	2018-08-24 15:07:27	heatwave热浪秋季女士纯色运动面料透气T恤半高领打底衫60203
206410634002175892	交易关闭	2018-08-24 15:01:01	heatwave热浪外套女拉链开衫带帽运动休闲上衣秋季修身60223，heatwave热浪秋季女士纯色运动面料透气T恤半高领打底衫60203

Item list (order detail list) dataê

The item list includes the order number and the products bought in a specific order. Product details such as name, price, amount, identification number, colors, etc are included. Each row only presents one item in an order. Timestamps are missing in the item list. There are fewer orders in the item list and the data is more incomplete.

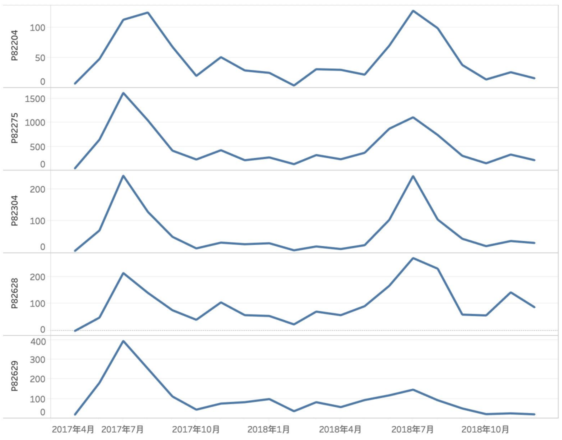


订单编号	标题	价格	购买数量	外部系统编号	商品属性
206677894478132351	heatwave热浪保守连体游泳衣女遮肚性感黑色显瘦聚拢短袖大码泳裙	518	1	82275	颜色分类：黑色;尺码：L
206727172469844374	heatwave热浪可爱甜美高弹印花游泳帽透气耐用不紧绷舒适不勒头	50	1	H319-7	颜色分类：花色
207310708587535922	heatwave热浪保守连体游泳衣女遮肚性感黑色显瘦聚拢短袖大码泳裙	518	1	82275	颜色分类：黑色;尺码：XL
207292060240853709	heatwave热浪保守连体游泳衣女遮肚性感黑色显瘦聚拢短袖大码泳裙	518	1	82275	颜色分类：黑色;尺码：XL
176101633762589196	热浪新款舒适时尚黑色显瘦遮肚性感泳装短袖保守平角连体游泳衣女	468	1	82769	颜色分类：黑色;尺码：L
207671855316541400	热浪新款舒适时尚黑色显瘦遮肚性感泳装短袖保守平角连体游泳衣女	468	1	82769	颜色分类：黑色;尺码：XL
206412454844175892	heatwave热浪秋季女士纯色运动面料透气T恤半高领打底衫60203	288	1	60203	尺码：L;颜色分类：蓝色

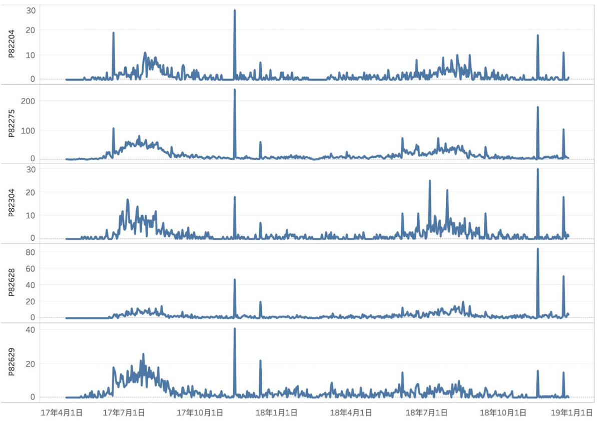
Integrated Data

Because many missing and incomplete data occurred during the integration, we chose five products that have a relatively more complete dataset. We look at the data in both daily and monthly scale to capture different patterns of data. The daily data starts from 2017-04-22 to 2018-12-18 and monthly data starts from 2017-05 to 2018-11.

Monthly Data

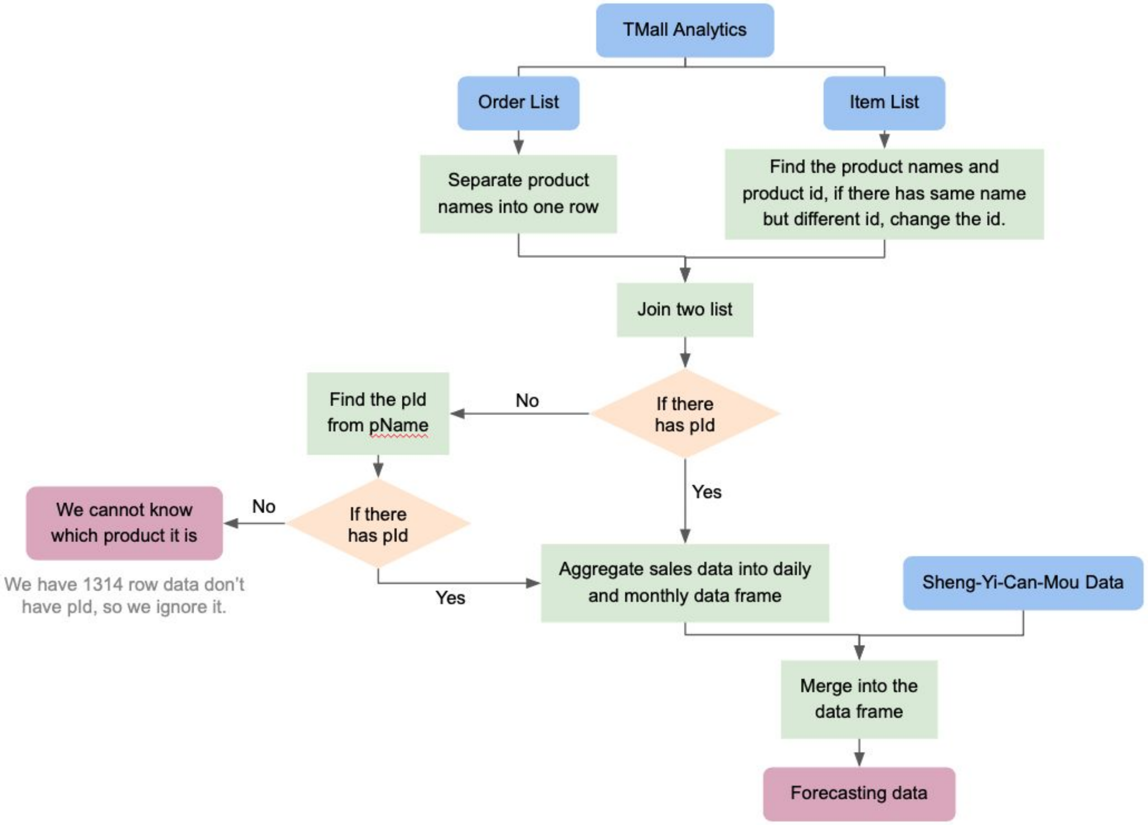


Daily Data



Preprocessing Method

First we deal with the order list and the item list. We try to match all product names and product identification number. We join two list but there still has many missing value. After preprocessing them, we miss 1314 records because we can not find the product identification number. Although we get many missing value, we still get some useful data and that’s helpful for our forecast.The following is the process of data preprocessing:

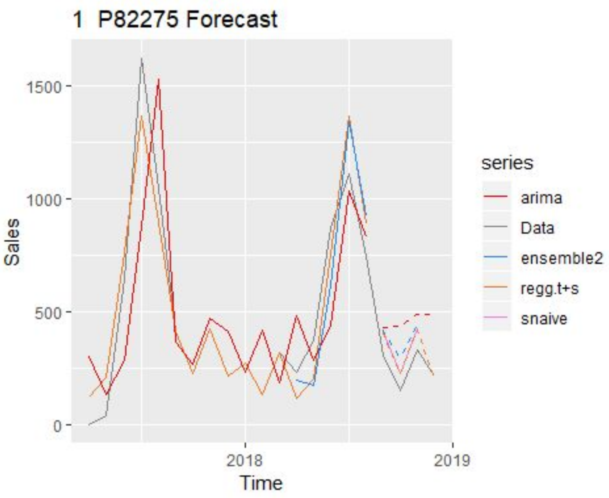

TMall Analytics
Order List
Item List
Separate product names into one row
Find the product names and product id, if there has same name but different id, change the id.
Join two list
If there has pId
No
Find the pId from pName
Yes
If there has pId
No
We cannot know which product it is
We have 1314 row data don’t have pId, so we ignore it.
Yes
Aggregate sales data into daily and monthly data frame
Sheng-Yi-Can-Mou Data
Merge into the data frame
Forecasting data

Forecasting Solutions using Monthly Data

Monthly Forecast

We use the data of the last three months as the validation set and others for training set. We tried some models which we show the RMSE results in the evaluation table below. We chose some models and plotted the results.

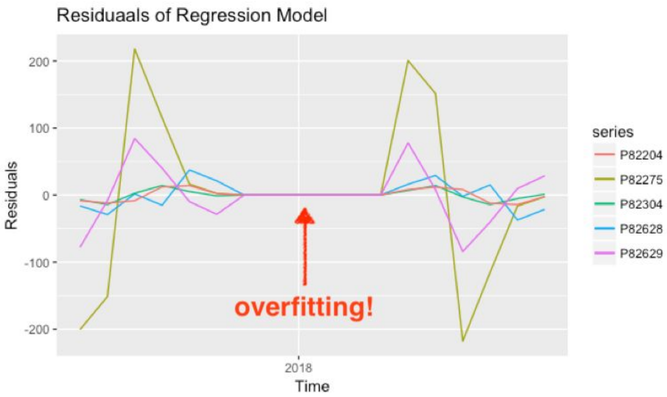
Monthly Forecast Result Plot



Evaluation

Monthly RMSE	Naive	sNaive (Benchmark)	ets	regression	arima	ensemble (snaive, regression, arima)
training	136.11	121.46	131.99	**40.13**	102.06	87.89
test	170.73	**38.88**	174.28	**41.64**	99.33	59.95

According to the forecast chart and RMSE table, we can infer that for the monthly data, the **sNaive** has the best performance. The regression model seems to perform well, but we can see it overfits in some months from the residuals plot below.



Moreover, the ensemble doesn't work because all models tend to over-forecast. Because of the shortage of data, the model is easy to overfit. Therefore, we recommend using daily data for forecast and avoid over forecasting the sales.