How busy will my restaurant be tomorrow? Forecasting the daily number of customers in each restaurant

Business Analytics Using Forecasting

Group 6

Edison Lee, Celia Chen, Sehyeon Jeong, Guan-Jie Chen, Web Yuan

2016/9~2017/1

國立清華大學 服務科學研究所
INSTITUTE OF SERVICE SCIENCE
NATIONAL TSING HUA UNIVERSITY

Executive Summary

Business Problem

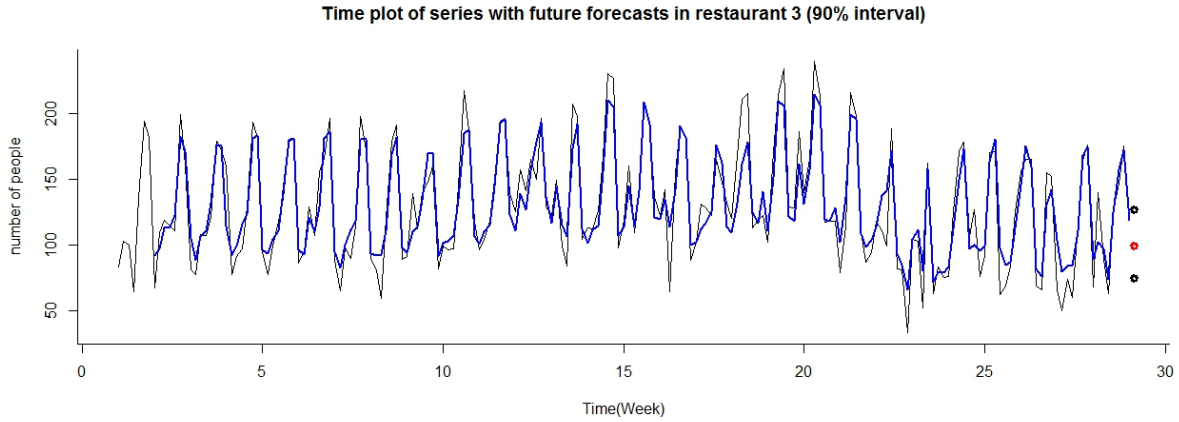
We are going to let manager of each restaurant know how busy they will be tomorrow by this business forecasting. The forecasted value would be used as a mental preparation of manager.

Data

The data is provided from iCHEF. It was taking the form of invoice which includes information of timestamp, items, and so on. And then, we make it into 2-column data (demonstrated in the right side of this paragraph) with preprocessing. We are measuring the daily number of people to forecast future number of people dine in the restaurant. Every restaurant has weekly tendency. Weekend has more customer than weekday. Some of restaurants show a slight decrease of people. The others have no specific decrease or increase.

restaurant 1	
timestamp	people
2016/7/12	6
2016/7/13	4
2016/7/14	49
2016/7/15	151
2016/7/16	180
2016/7/17	190
2016/7/18	144
2016/7/19	137
2016/7/20	123
2016/7/21	163

Forecasting Solution & Limitation



✧ Sample of one restaurant's forecasting solution

- **Forecasting Solution:** The manager will receive the result of this forecast one day ahead. They would have mental preparation for tomorrow, and prepare next day job allocation with our forecasted value.
- **Forecasting Limitation:** The lack of data amount harms the accuracy of future forecasting. Because of it, if we provide to manager "Under-forecasted value", we might let manager mishandle the job allocation.

Recommendation

To complete this project, we tried to choose the different number of days forecasting. We found the performance of 2-day and 3-day ahead are not bad. Based on this result, we recommend not to choose only 1-day ahead forecasting so that the manager can get the forecast earlier. In addition, we highly recommend that iCHEF provides data on daily basis for the suitability of forecasting.

Detailed report

Problem description

- **Business goal:** Let manager of each restaurant know how busy they will be tomorrow.
- **Description:**
  - ❏ **Client:** manager of the restaurant
  - ❏ **Stakeholders:** iChef, Restaurants (owner, staff, customer)
  - ❏ **Benefit:** the manager would have mental preparation for tomorrow, and prepare next day job arrangement with our forecasting result.
  - ❏ **Opportunity:** some job could be done when the number of customer is low, so the staff would provide more efficient service.
  - ❏ **Shortcoming:** under-forecasting result might let staff unable to handle the job arranged by manager.
- **What would be considered a success?**

  We set the same day number of customer last day as benchmark, if the forecasting result is more accurate than it, we consider the result is a success forecast.
- **Forecasting goal:** Forecasting the daily number of customers in each restaurant.
- **Description of forecasting object:**
  - **Number of Series:** (5 restaurants) * (Daily # of dine-in customers) = 5 series
  - It is a forward-looking (prospective) task.
  - **Forecasting horizon (k):** one-day-ahead.
  - **Time period (t):** daily
  - **Value of the series at time t (y_t):** Daily number of dine-in customers in each restaurant
- **How will the forecasts be used?**
  - ❏ The forecast will be used as a mental preparation of how busy will be tomorrow. On top of knowing the approximate number of customer tomorrow, the manager can try to arrange the job of the day.

Data description

- **Source:** iCHEF
- **Measure:** SUM(people), as. Date(timestamp)
- **Time period:**
  From 04/01/2016 to 10/31/2016: 2 restaurants - 7months
  From 05/04/2016 to 10/31/2016: 1 restaurant - 6months
  From 05/05/2016 to 10/31/2016: 1 restaurant - 6months
  From 07/18/2016 to 10/31/2016: 1 restaurant - 3months and a half
  (5 restaurants) * (Daily # of dine-in customers) = 5 series
- **Frequency of collecting data:** Daily
- **Characteristic of the series:**
  - ❏ See all restaurants' raw time series in appendix 1
  - ❏ Each of the series has weekly seasonality.
  - ❏ Some of the series have slightly downward trend. Some of them have no trend.

restaurant 1		restaurant 2		restaurant 3		restaurant 4		restaurant 5	
timestamp	people	timestamp	people	timestamp	people	timestamp	people	timestamp	people
2016/7/12	6	2016/4/11	105	2016/4/11	83	2016/5/16	58	2016/5/16	312
2016/7/13	4	2016/4/12	128	2016/4/12	103	2016/5/17	60	2016/5/17	231
2016/7/14	49	2016/4/13	115	2016/4/13	100	2016/5/18	69	2016/5/18	287
2016/7/15	151	2016/4/14	119	2016/4/14	64	2016/5/19	74	2016/5/19	276
2016/7/16	180	2016/4/15	148	2016/4/15	135	2016/5/20	100	2016/5/20	310
2016/7/17	190	2016/4/16	196	2016/4/16	194	2016/5/21	201	2016/5/21	408
2016/7/18	144	2016/4/17	214	2016/4/17	182	2016/5/22	216	2016/5/22	411
2016/7/19	137	2016/4/18	81	2016/4/18	67	2016/5/23	69	2016/5/23	288
2016/7/20	123	2016/4/19	103	2016/4/19	109	2016/5/24	59	2016/5/24	222
2016/7/21	163	2016/4/20	137	2016/4/20	119	2016/5/25	66	2016/5/25	212

✧ Sample of first ten rows for five series

Brief data preparation details

- Description of raw data: This file contains all items in every invoices across five restaurants. Each raw stands for an item in one specific invoice.

nvoice_uuid	item_name	item_uuid	people	type	outset	price	timestamp	timestamp	restaurant_uuid	sales_amount
00026EA-B2E6-41C8-9A99-6C52E825FE4F	羅馬-松露野菇	48a0422a-9bf0-4fe2-ba6e-768e6c6e239c	1	combo	takeout	380	2016-06-15 09:27:29.789692	2016-06-15 09:27:29.789692	6d0ebab3-edf8-4e04-a947-1973e76ab11f	1140
00026EA-B2E6-41C8-9A99-6C52E825FE4F	拿-經典辣味燻雞	3090bcb2-16e9-4971-8087-b3f92ee6343d	1	combo	takeout	220	2016-06-15 09:27:29.789692	2016-06-15 09:27:29.789692	6d0ebab3-edf8-4e04-a947-1973e76ab11f	1140
00026EA-B2E6-41C8-9A99-6C52E825FE4F	拿-經典瑪格	5ebb6673-0086-4b8a-a052-cdb3424ee3c3	1	combo	takeout	180	2016-06-15 09:27:29.789692	2016-06-15 09:27:29.789692	6d0ebab3-edf8-4e04-a947-1973e76ab11f	1140
00026EA-B2E6-41C8-9A99-6C52E825FE4F	電話	6785d383-5a26-4133-ae11-1e9c1bd4462b	1	item	takeout	0	2016-06-15 09:27:29.789692	2016-06-15 09:27:29.789692	6d0ebab3-edf8-4e04-a947-1973e76ab11f	1140
00026EA-B2E6-41C8-9A99-6C52E825FE4F	羅馬-義式果香燻雞	9ba372f0-038f-4b3b-9afa-833abe6bdfe0	1	combo	takeout	360	2016-06-15 09:27:29.789692	2016-06-15 09:27:29.789692	6d0ebab3-edf8-4e04-a947-1973e76ab11f	1140
00026EA-B2E6-41C8-9A99-6C52E825FE4F	拿-堤諾先生	57525b6c-4086-4bc1-afa4-aa3def92eab4	1	combo	takeout	200	2016-06-15 09:27:29.789692	2016-06-15 09:27:29.789692	6d0ebab3-edf8-4e04-a947-1973e76ab11f	1140
00026EA-B2E6-41C8-9A99-6C52E825FE4F	羅馬-西西里蒜味海鮮	76381357-b595-4836-a35a-013da575d847	1	combo	takeout	400	2016-06-15 09:27:29.789692	2016-06-15 09:27:29.789692	6d0ebab3-edf8-4e04-a947-1973e76ab11f	1140
00037B7-5F4E-4196-867B-6D0890E03E8F	酥炸酒醋蘑菇	f703cbf4-5c90-4958-b6e1-a1d8aa97723f	2	combo	indoor	99	2016-05-13 13:44:20.797426	2016-05-13 13:44:20.797426	535b23c0-728f-4ced-8ad6-c8ecd8ae379d	835
00037B7-5F4E-4196-867B-6D0890E03E8F	羅馬-堤諾先生	88a36403-6c9b-4beb-b1b9-7ea40463061e	2	combo	indoor	320	2016-05-13 13:44:20.797426	2016-05-13 13:44:20.797426	535b23c0-728f-4ced-8ad6-c8ecd8ae379d	835

✧ Sample of raw data file

- **Pre-process steps:**
  - ❑ Step 1: Aggregate item-level data to invoice-level data, so that we can have the data with each row representing for one invoice.
  - ❑ Step 2: Filter out only dine-in invoices and then separate to five data files by restaurant. Each file includes all invoices in one restaurant.
  - ❑ Step 3: For each file, aggregate invoice-level data to daily-level.
  - ❑ Step 4: Because we find that there are "NEW opening days" for each series (restaurant), we remove approximately one to two weeks.
  - ❑ Step 5: We find some missing values in our series. We impute those missing values with last week value.

Forecasting solution

- **Evaluation:** MAE, MAPE, RMSE

  The purpose of this forecast is to let the manager know the approximate number of customer tomorrow, so that he/she can know how busy they will be for the next day. Thus, the manager will receive the result of this forecast as an interval. Also, the more precise the forecast is, the better result it will be. The evaluation of models, therefore, will be the metrics (namely MAE, RMSE, MAPE).
- **Methods applied and evaluation:**
  - ❑ **Benchmark:** Seasonal Naive
  - ❑ All of the forecast are conducted using software R and Excel.
  - ❑ See the chart of all restaurants' evaluation in appendix 2
- **Model building:**

  See each model we try in all restaurants comparing with seasonal naïve in appendix 3. All restaurants have a brief description about how each model perform and what model we choose in the end.

Time plot of series with future forecasts for each of the 5 series

Since the last data point in our data is October 31, 2016 for all five series, we provide one-day ahead future forecast (November 1, 2016) in appendix 4. We also provide 90% prediction interval with point forecasts, and we use empirical way to compute our prediction interval. The 90% prediction interval is conducted by using the 5th and 95th percentiles from the forecast errors within validation periods.

Conclusions

- **Advantages:**

Help managers to better arrange daily job to staff.

- **Limitations:**
  - ❏ The longest time period in our data is only seven-month, which makes us hard to know if there are yearly seasonality or not.
  - ❏ For now, the forecast will be less valuable since the size of people is too small and the range forecast error is wide due to the small amount of data. Take restaurant 4 for example, the forecast interval of future is (0,75) with the 90% CI (-32, +44); however, the number of people in weekday is about 30-70 in average. Thus, it seems that it doesn't benefit much by doing this analysis now. We think the key reason is that the distribution of error in validation period is not stable enough now. More data should be added to let the prediction interval smaller so that the forecasting result would be more accurate. As the result, the distribution of error in validation period would become more stable since we could collect more data and error in the future.
- **Recommendations:**
  - ❏ **For the client:**

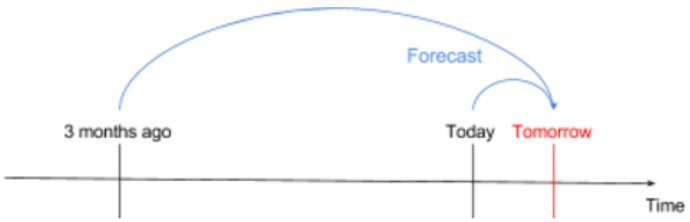
We had done 7 days ahead forecast for one restaurant. We found the performance of 2 and 3 three days ahead is not so bad. As the result, we can try 3 or more days ahead forecast in the future. The client can get the data earlier, so its forecast result might be more valuable.

  - ❏ **For the people who want to continue developing this forecast:**

We tried one external information(weekday/weekend) in our forecasting model, but it only one restaurant performs well. It can be added more external information so that the forecasting result would be better.

  - ❏ **For data collecting frequency:**

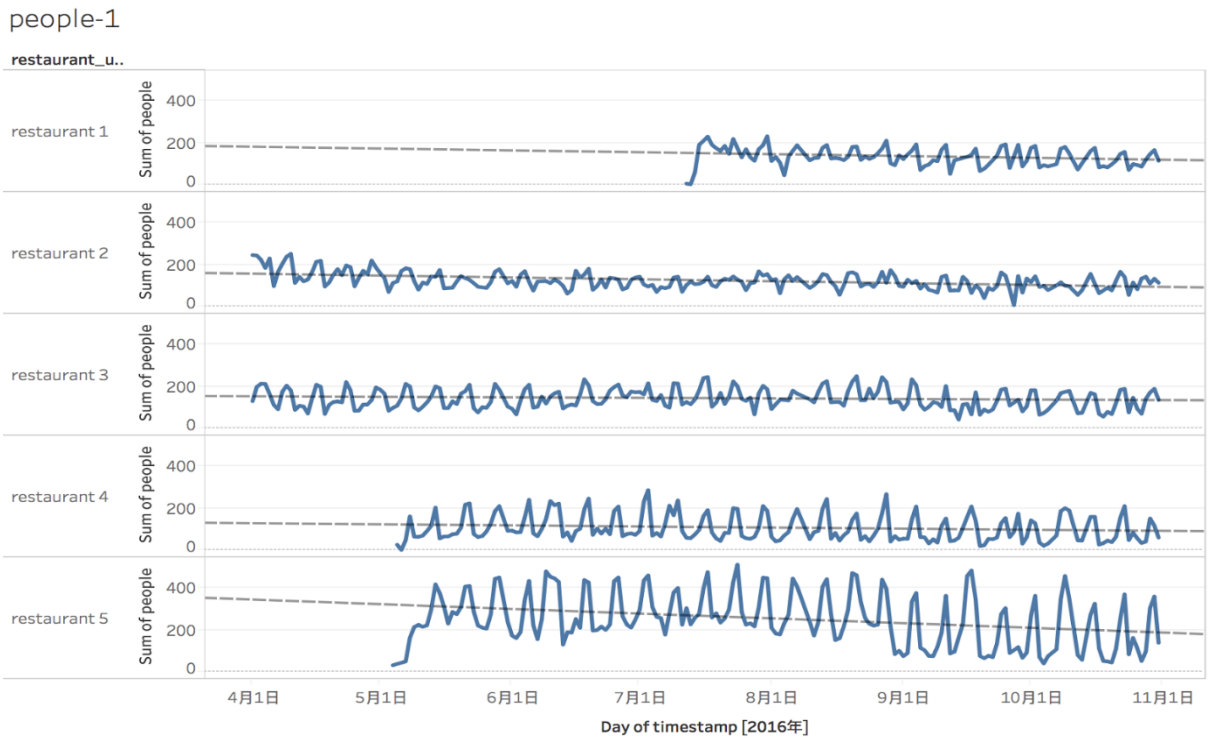
In this project, we assume that we could get data daily compare to the situation now that we have three-month delay to get data. Take the plot below for example, if we want to forecast tomorrow, it would be more suitable to use today's data than use 3 months ago data. Therefore, we highly recommend that iCHEF could provide data on daily basis.



✧ Data collecting frequency

Appendix

1. All restaurants' raw time series



The trend of sum of people for timestamp Day broken down by restaurant_uuid. The view is filtered on restaurant_uuid, which keeps restaurant 5, restaurant 1, restaurant 3, restaurant 4 and restaurant 2.

2. All restaurants' evaluation

Restaurant	Split	Evaluation	SNaive	Smoothing	Regression	NeuralNetwork	Ensemble
1	Training	MAE	17.33803	13.80804	13.79573	13.73271	13.46329478
Edison		RMSE	23.92741	18.84266	19.15401	18.15823	18.61714
		MAPE	17.0555	14.0667	14.11471	13.7287	13.83849564
	Valid	MAE	16.14286	12.71003	10.53106	16.58203	11.27310794
		RMSE	25.72103	17.44983	16.13065	22.07386	16.22562037
		MAPE	17.43884	13.87509	10.42628	18.5624	11.87447964
2	Training	MAE	25.50296	19.21825	19.49779	2.279964	10.28832
Celia		RMSE	32.03511	24.24228	24.91885	3.198848	13.19963
		MAPE	33.4718364	29.677394	30.33776433	4.802939099	16.88788197
	Valid	MAE	25.57143	19.98456	19.58672	23.97424	20.85494
		RMSE	33.90112	22.18147	22.85296	28.60689	23.71199
		MAPE	28.7794	22.75178	21.3125	30.27472	25.21227
3	Training	MAE	26.39506173	23.97259	30.41537	12.13355	18.20785
		RMSE	34.82212	32.141	38.52742	15.5222	23.89107
Lia		MAPE	23.51152	21.1508	26.33493	11.27261	15.00124
	Valid	MAE	31.9111715	41.5032	46.56517	31.07775	31.91662865
		RMSE	51.97561	50.31712	51.37403	38.95223	42.68426
		MAPE	35.86899	46.2614	43.83687	35.58824	28.36294
4	Training	MAE	28.53	21.03	19.73	12.52	17.3
		RMSE	72.81	52.04	50.05	28.67	42.41
Web		MAPE	32.85	27.51	24.28	17.01	22.4
	Valid	MAE	36.64	25.3	21.86	38.39	25.97
		RMSE	103.38	66.03	62.64	98.5	67.76
		MAPE	84.89	41.96	37.64	100.2	55.56
5	Training	MAE	53.53636	25.743842	39.6224	4.131969	24.25
		RMSE	75.76369	18.4267	56.10384	5.848232	41.88
		MAPE	25.36274	24.102744	32.94861	2.699733	18.17
Jack	Valid	MAE	51.53571	32.52478	47.55487	79.570422	35.84
		RMSE	86.81569	26.57346	78.2443	63.30349	28.32
		MAPE	32.20483	28.42743	46.5784	68.978873	41.34

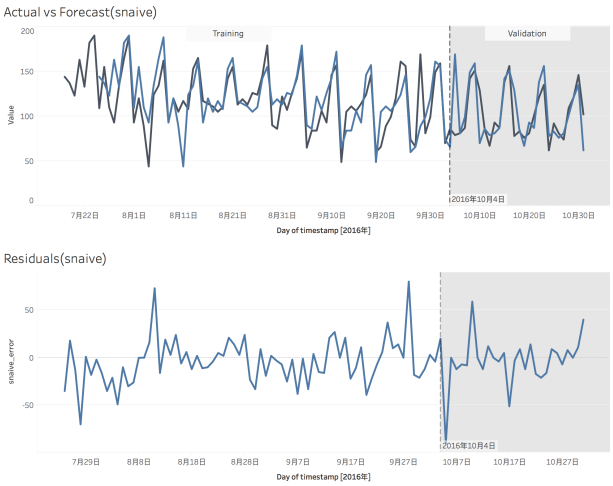
3. Model building

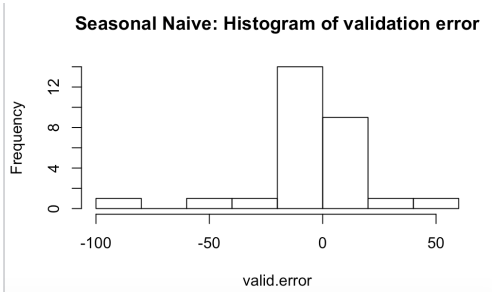
- **Restaurant 1:**

In addition to evaluation metrics, we can have more clear understanding from both time plots and histogram of error distribution. Therefore, from the following five histograms, we find that all of our five models for restaurant 1 have higher probability to be over-forecast.

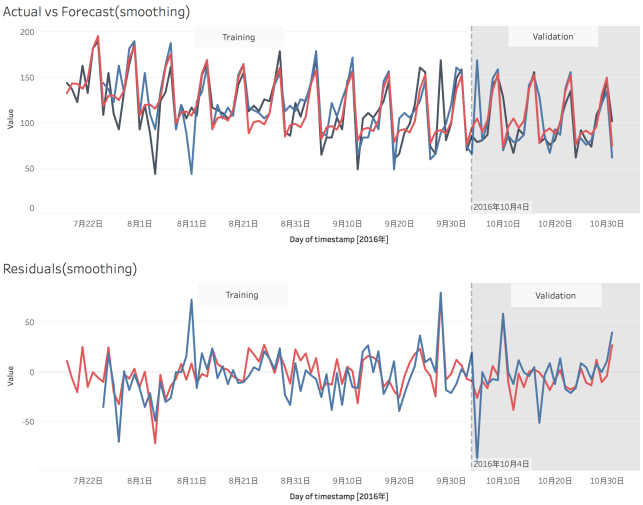
According to time plot of residuals for smoothing method, the forecast errors seem to be more stable than other methods. (However, we don't select this for our best model due to evaluation metrics.) Based on evaluation metrics, we choose regression model for the best model for restaurant 1. Furthermore, according to histogram of error distribution of regression model, we can easily find out that about 67% of forecast errors are between -10 to 10. It makes regression model perform better than other models.

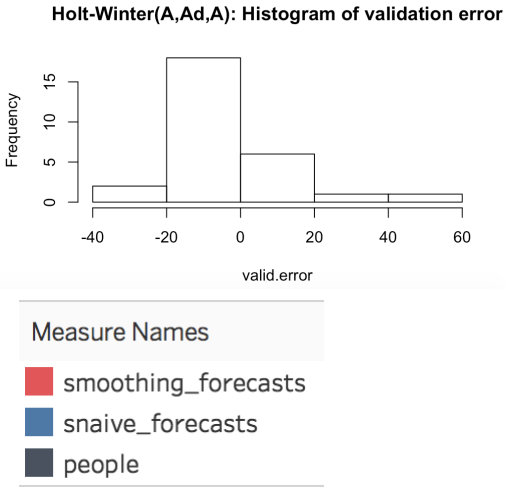
1. Seasonal Naive:



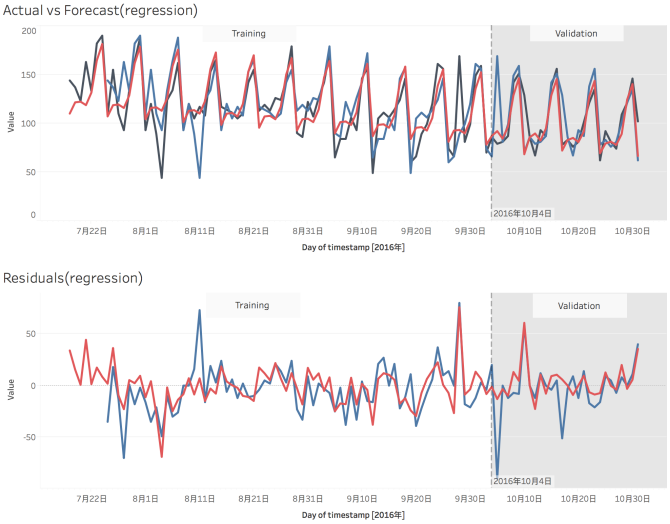


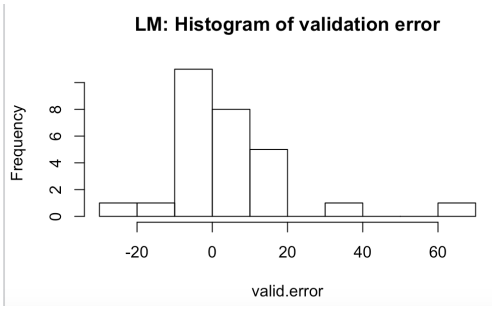
2. Smoothing



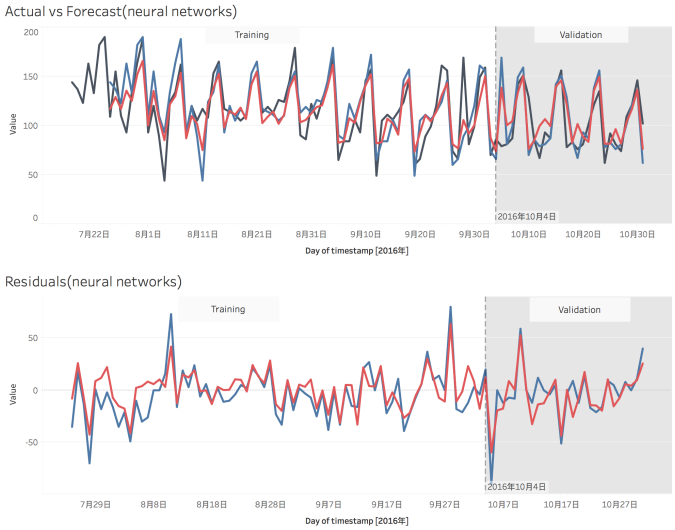


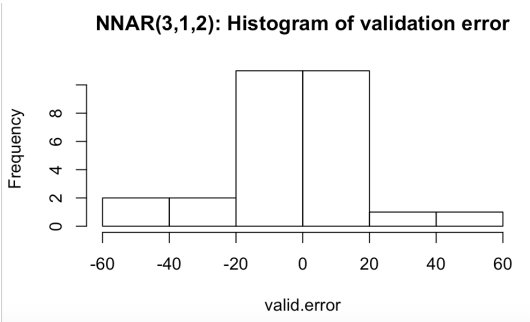
3. Regression



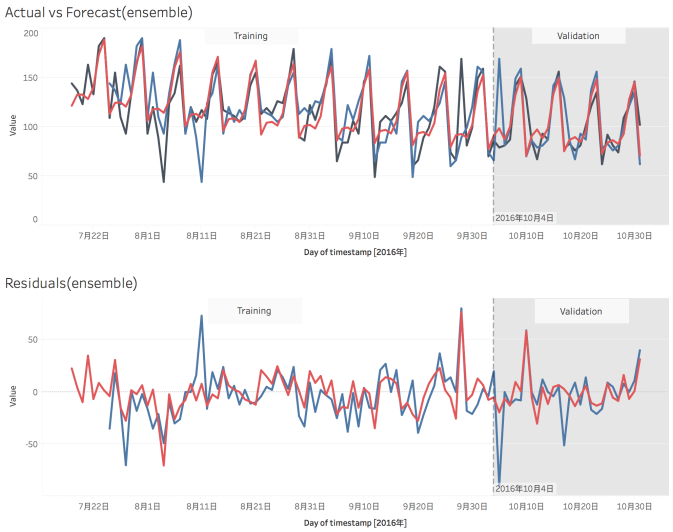


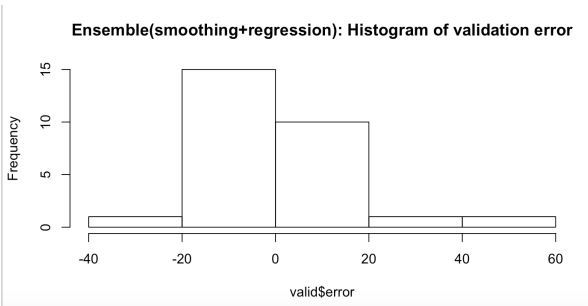
4. Neural Network





5. Ensemble (smoothing + regression)

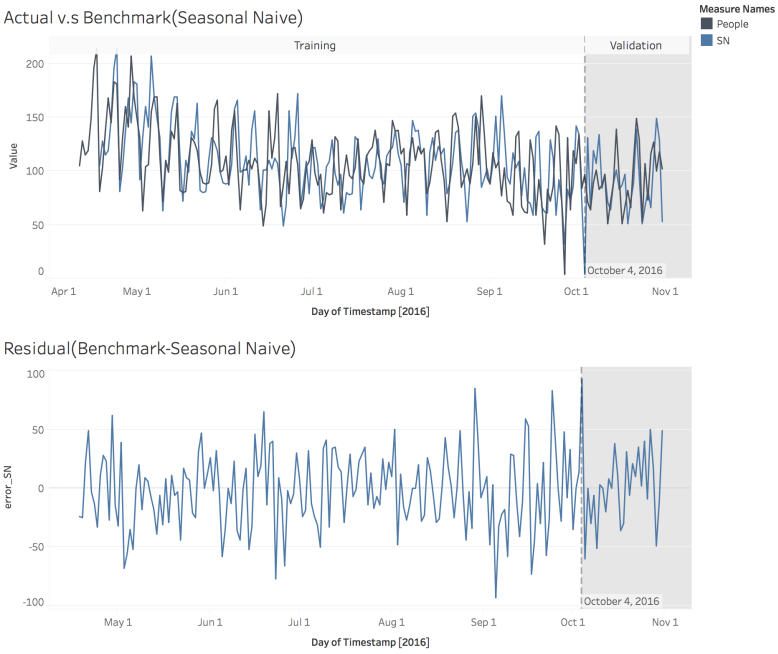


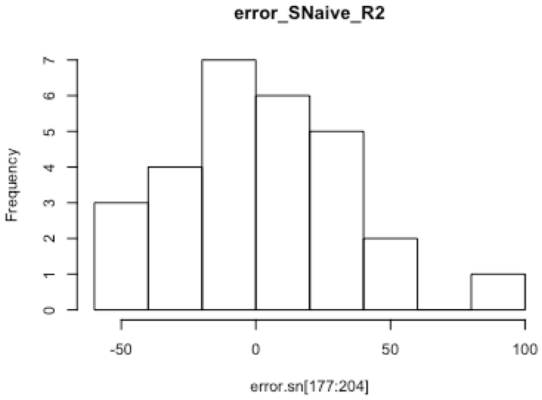


- **Restaurant 2:**

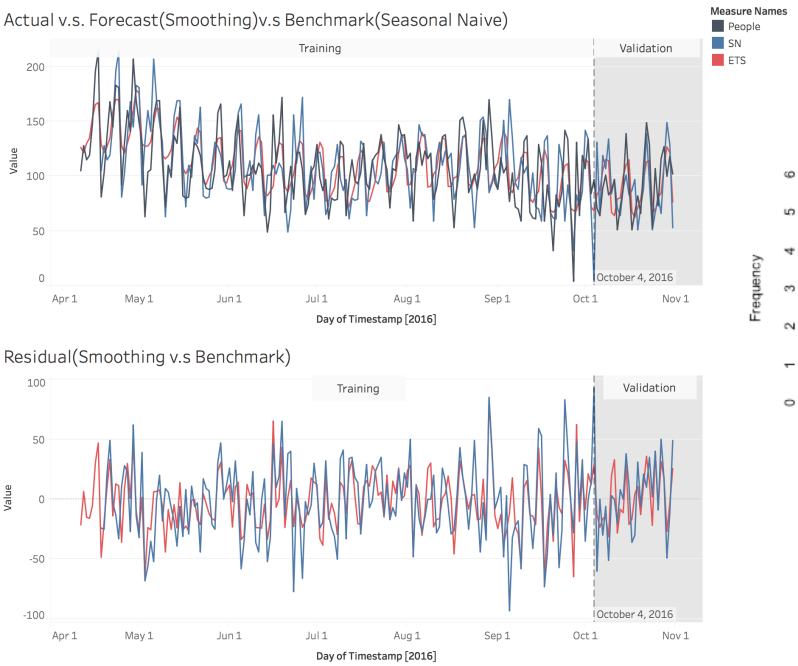
Restaurant 2 has a sight downward trend, a weekly seasonality as the other series and a crazier noise. Because of the crazy noise, it is hard to catch the pattern and forecast well. According to the performance metrics of 5 methods, the best model is Smoothing. Although Smoothing is not good enough, it beats the performance of the benchmark (Seasonal Naive). If we want to forecast this series better, we might want to try differencing or ARIMA to catch more information.

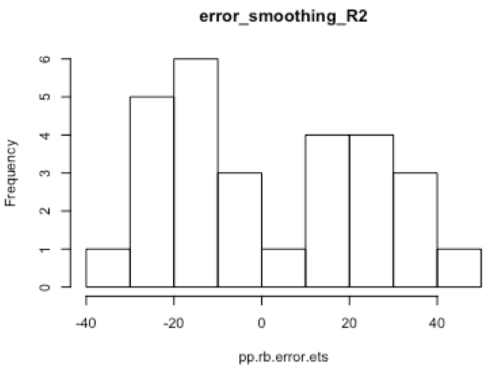
1. Seasonal naïve:



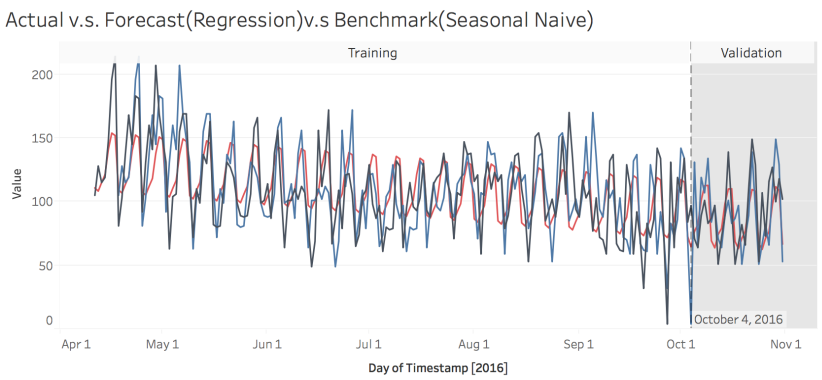


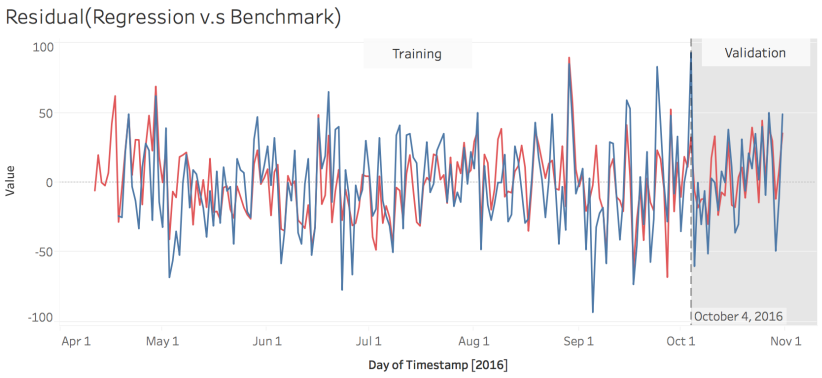
2. Smoothing

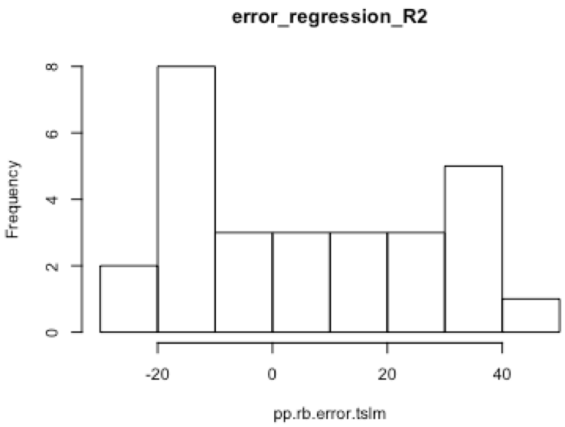




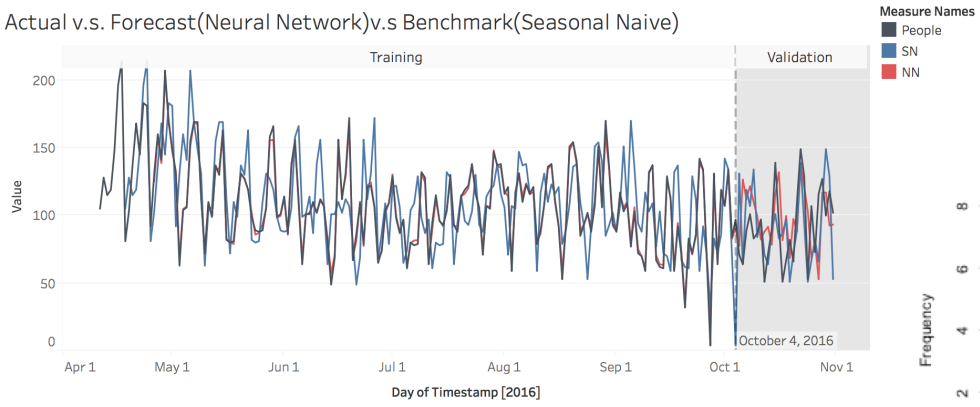
3. Regression

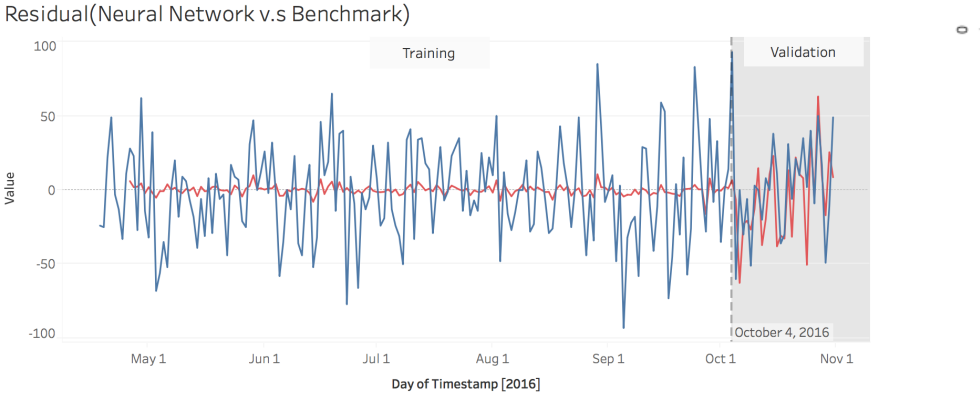


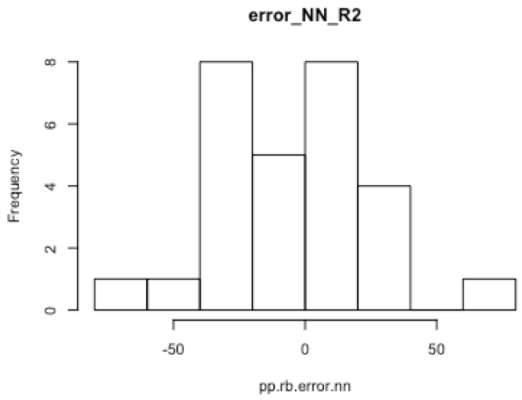




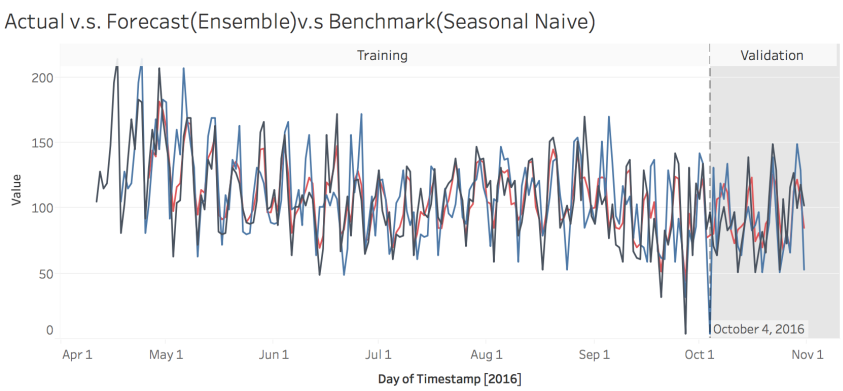
4. Neural Network

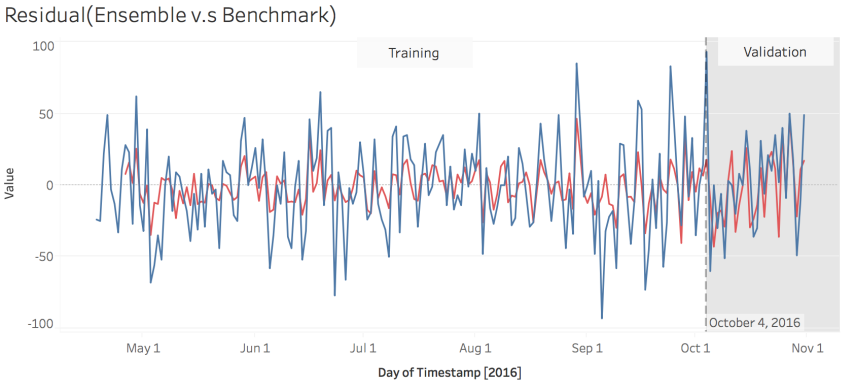


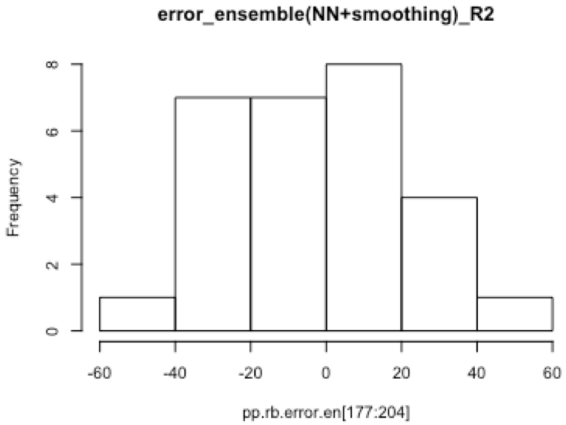




5. Ensemble (neural + smoothing)



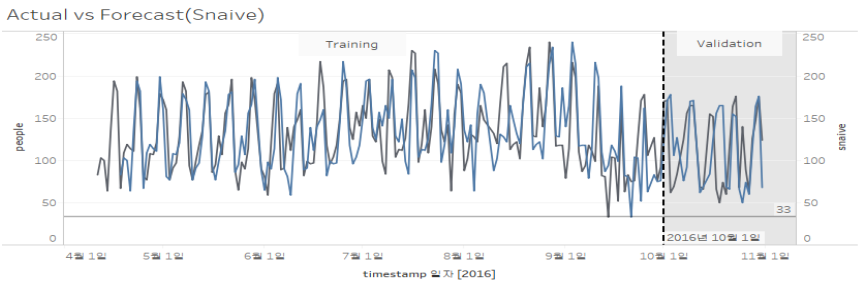


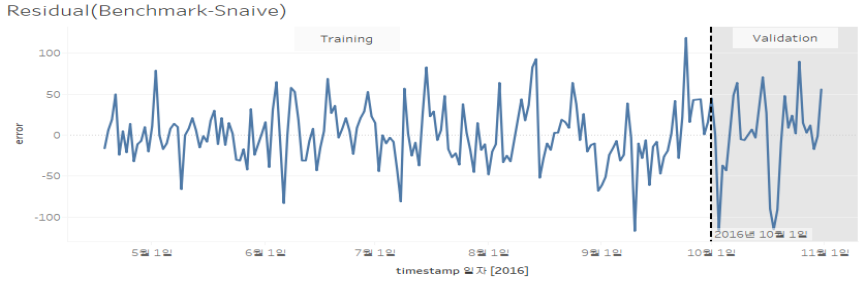


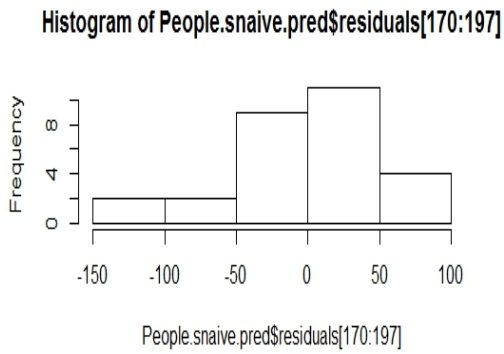
- **Restaurant 3:**

Restaurant 3 has lots of customers compared with others. It leads the forecasting result to have bigger residuals. Thus, if we don't care over-forecasting or under-forecasting, we would recommend Neural Network or Ensemble (Neural Network + Seasonal Naive) model to get most accurate forecasted value by the evaluation metrics.

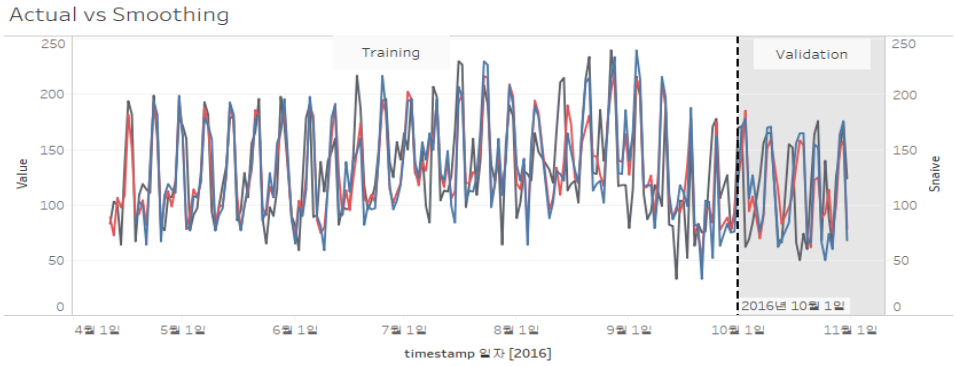
1. Seasonal naive

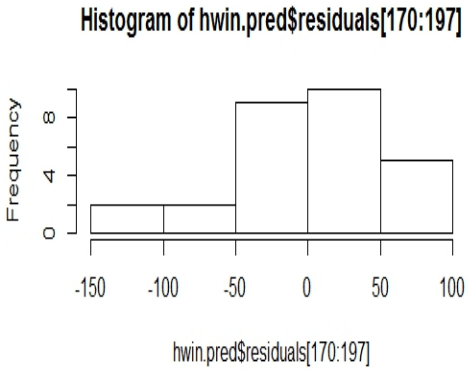


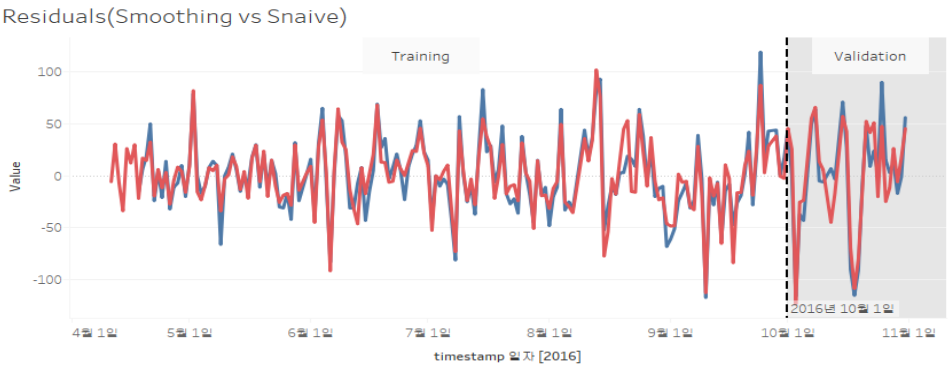




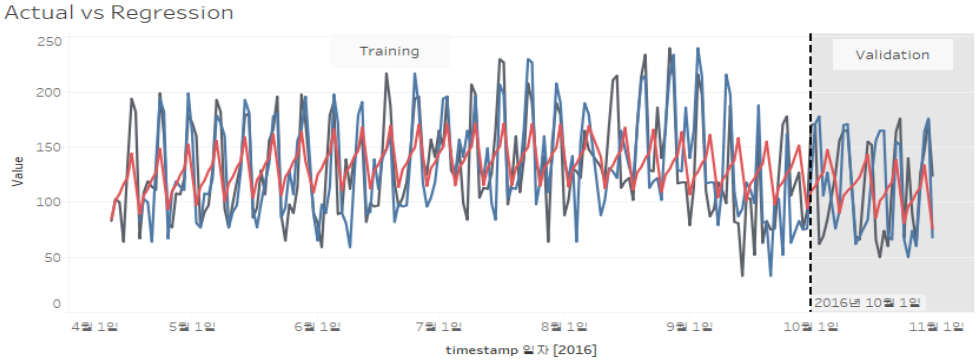
2. Smoothing

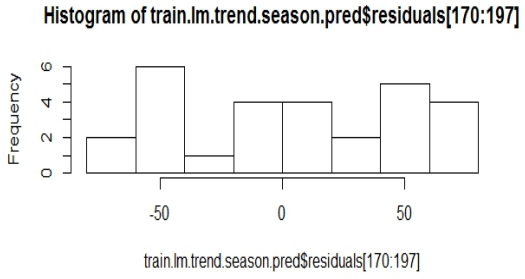


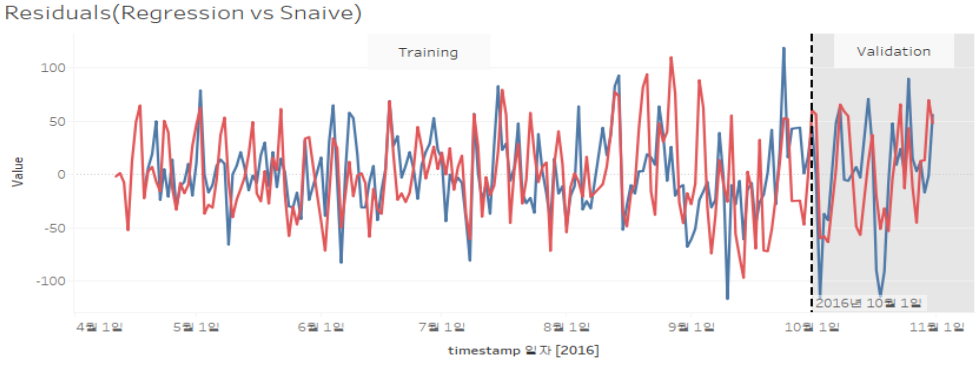




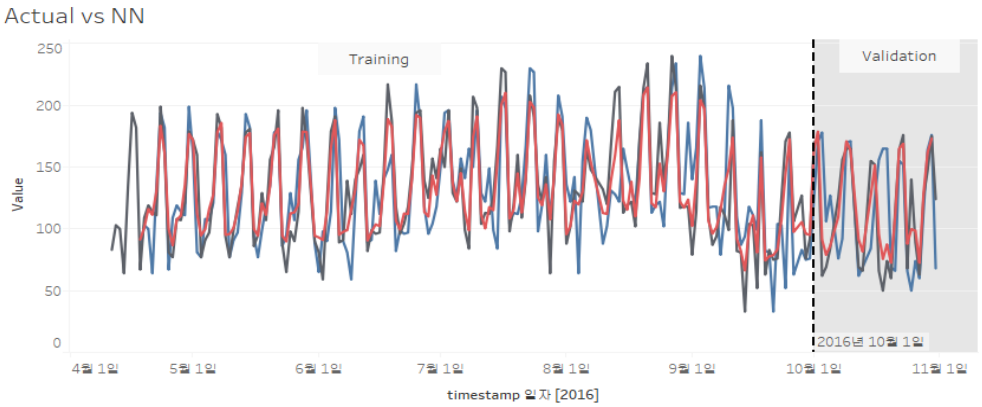
3. Regression

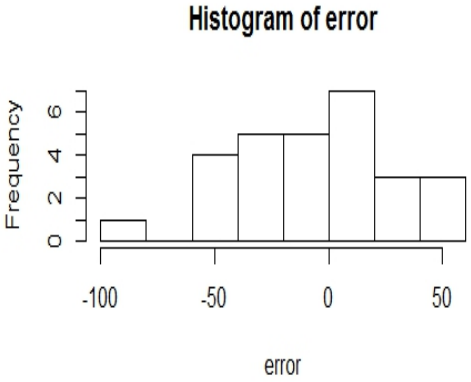


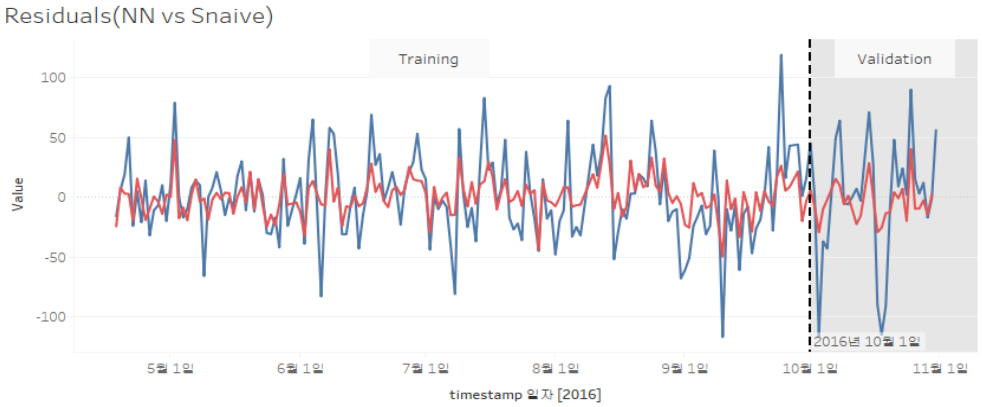




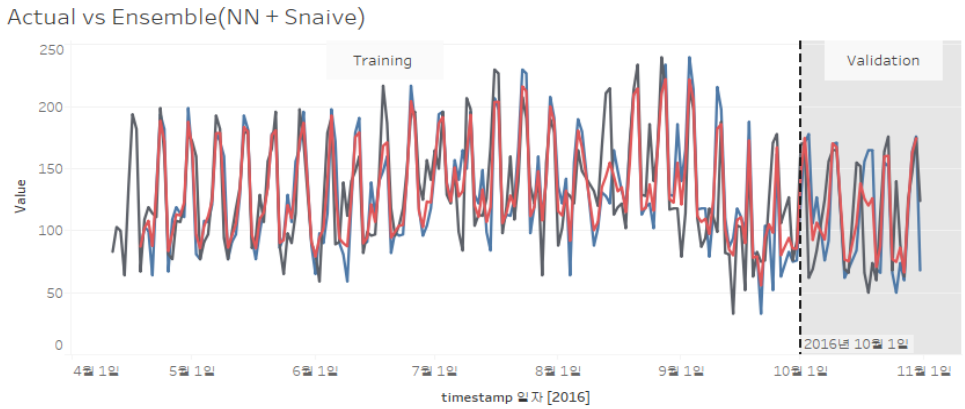
4. Neural Network

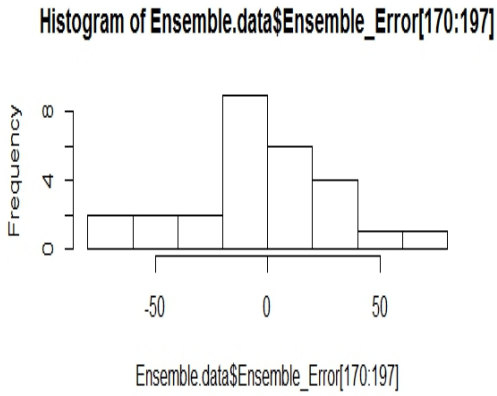


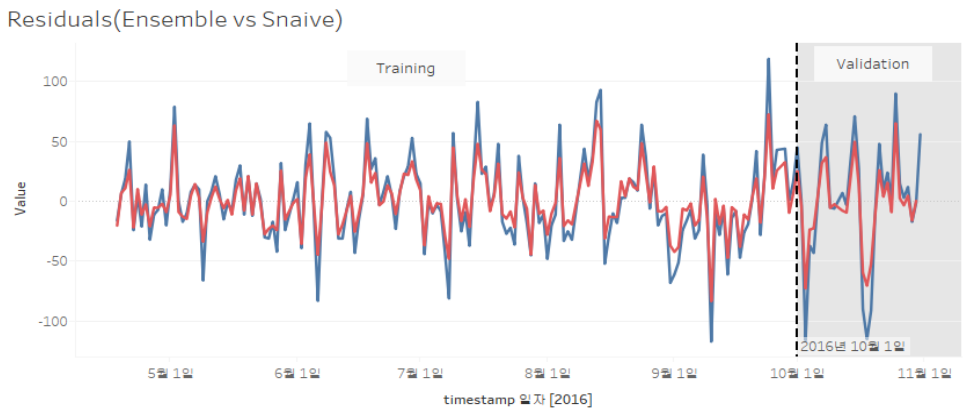




5. Ensemble (seasonal naïve + neural network)



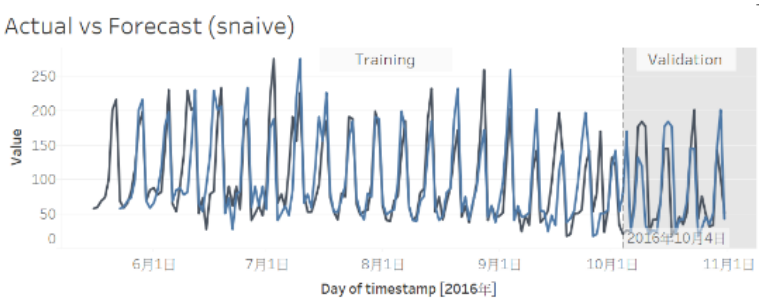


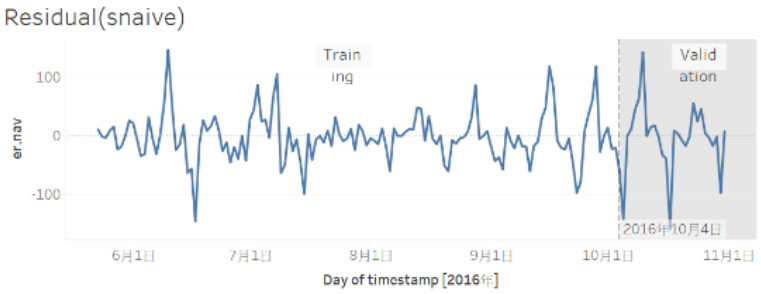


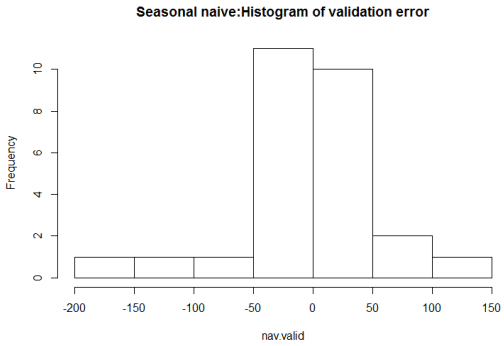
- **Restaurant 4:**

In restaurant 4, we can see that its time series have stable seasonality, but its trend is not clear. Besides that, we can find that neural network might be overfitting so that the residual of validation period is much more bigger than other model. Instead of neural network, all other model's distribution of residual show that most errors are under-forecasting, so we just don't care whether it is over-forecasting or under-forecasting. We also find there is nothing improvement in ensemble model, so I choose regression as the best model by choosing the lowest MAE, MAPE and RMSE model with our evaluation metrics.

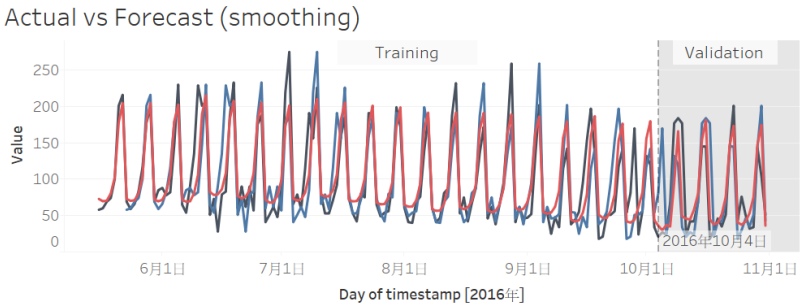
1. Seasonal naïve

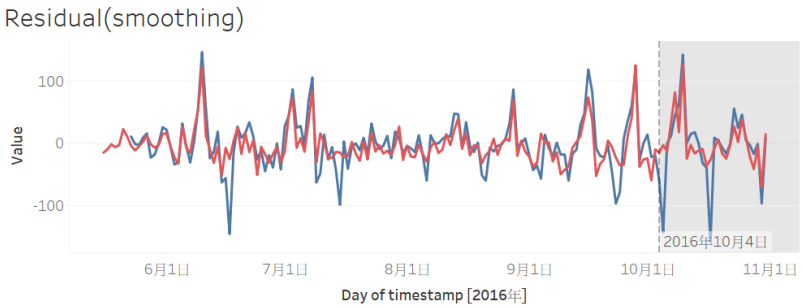


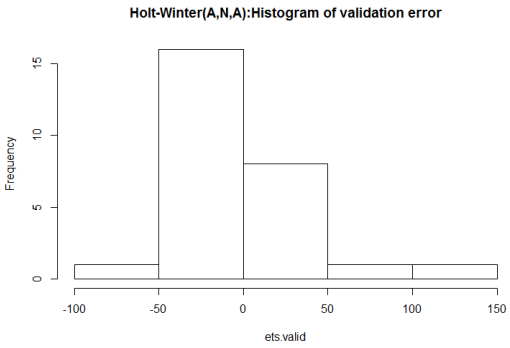


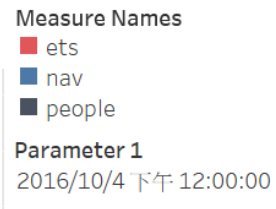


2. Smoothing

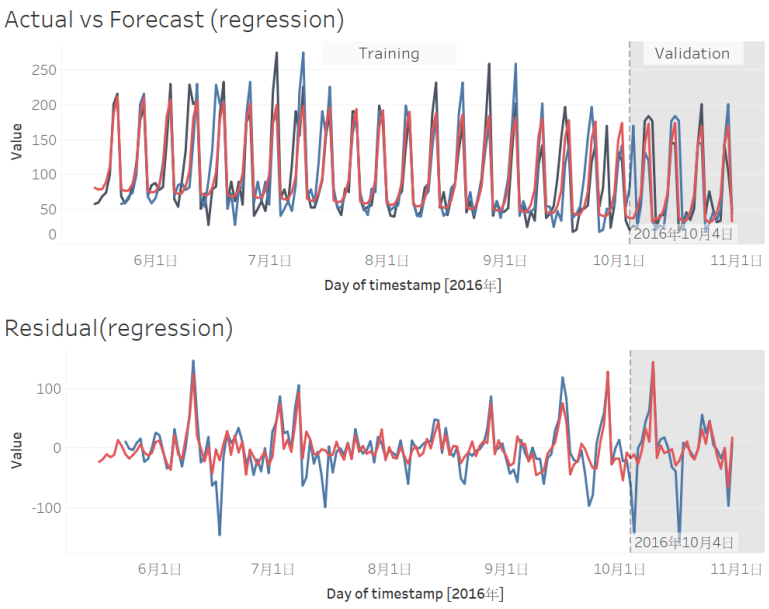


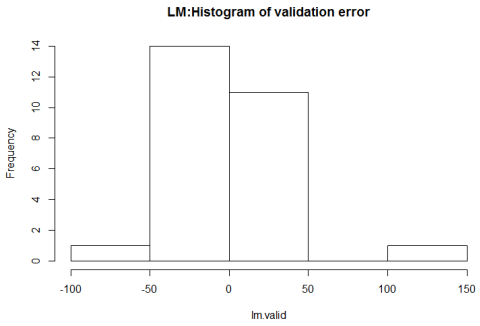




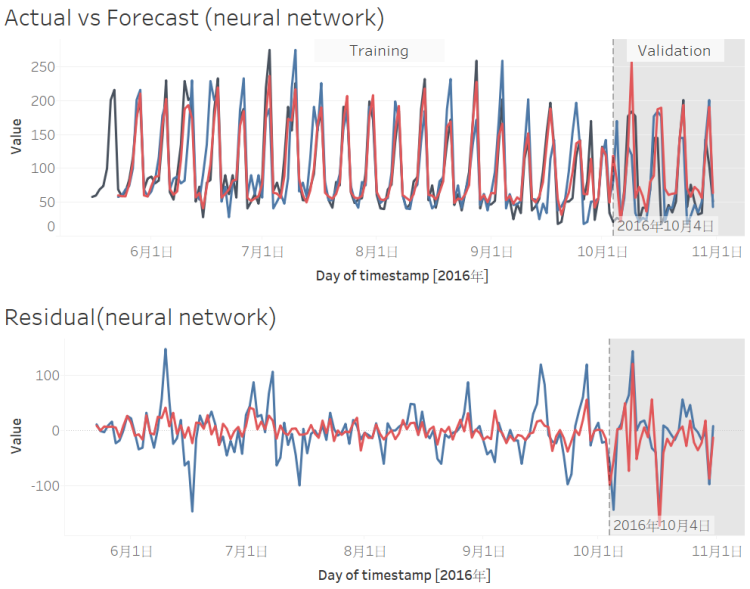


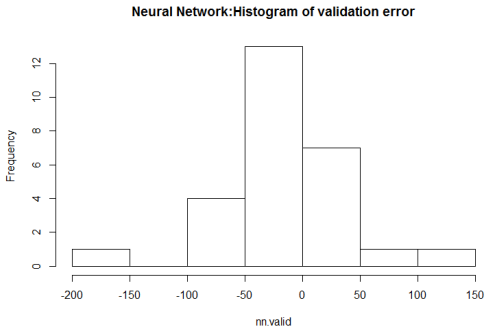
3. Regression



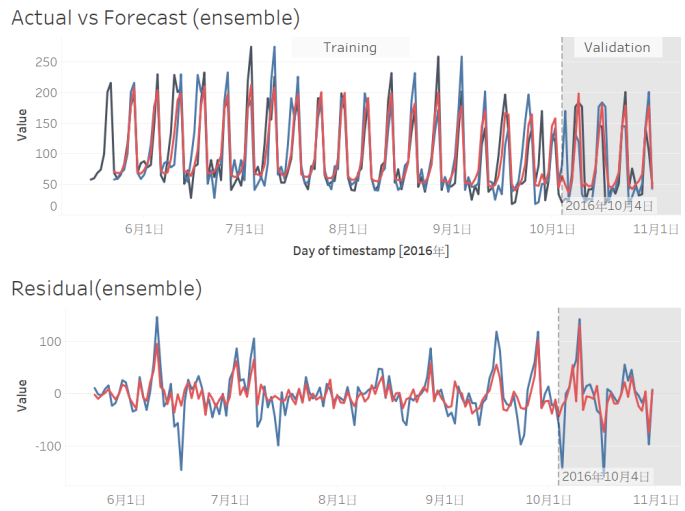


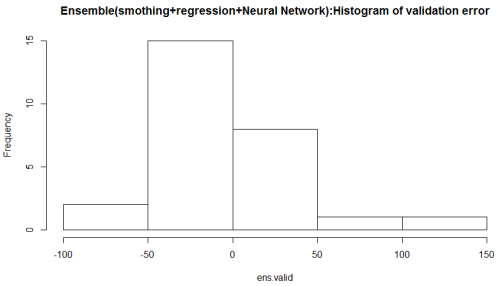
4. Neural Network





5. Ensemble (seasonal naïve + regression +neural network)

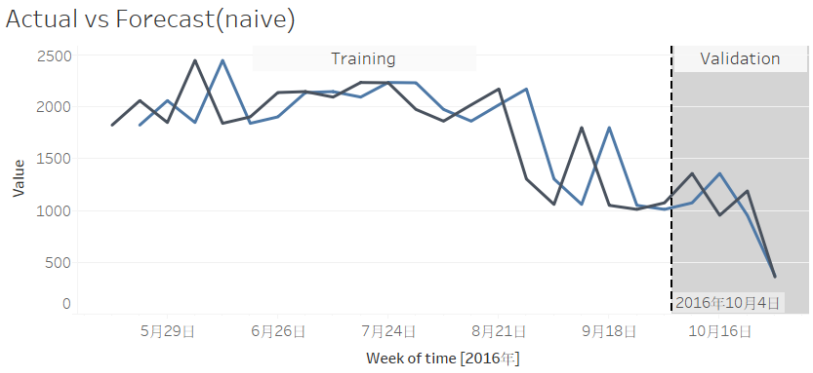


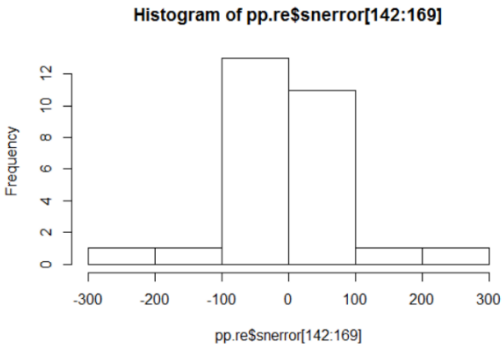


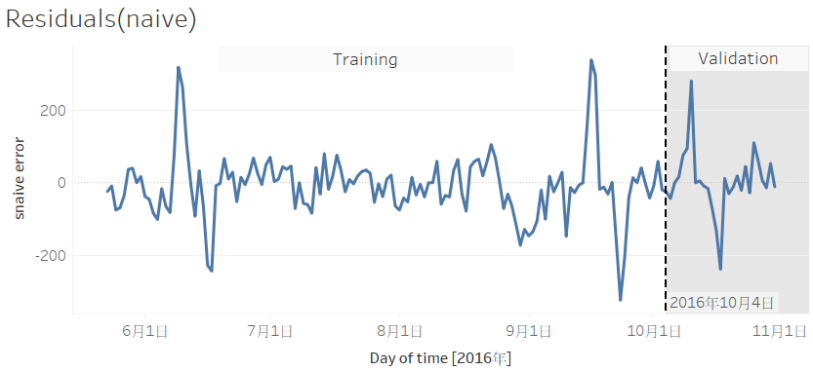
- Restaurant 5:

In restaurant 5, judging from the time plot, we can see that all model performs quite well in the training period. However, in the validation period, neural network performs the worst in the validation period (even though it performs the best in training period), so we do not choose the neural network model. And then, we have done the metrics (MAPE, RMSE, and so on) to compare the left three model, and finally found that smoothing performs better (in the histogram plot we can also see that most errors are in the range of -50 to 50, which makes its performance better).

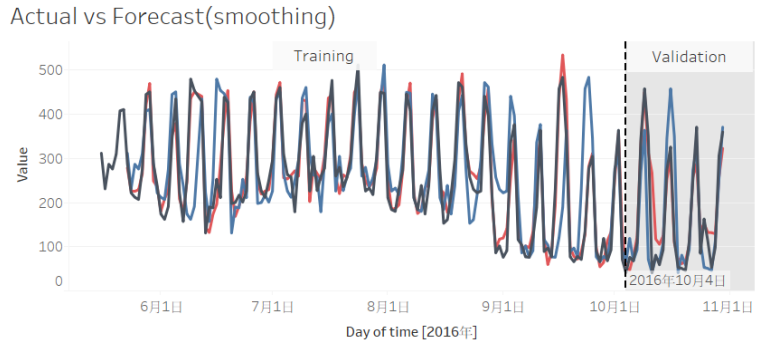
1. Seasonal naive

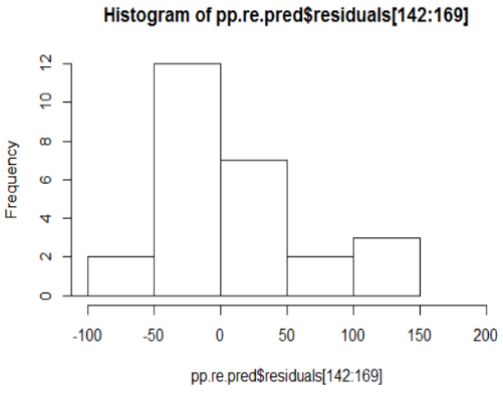


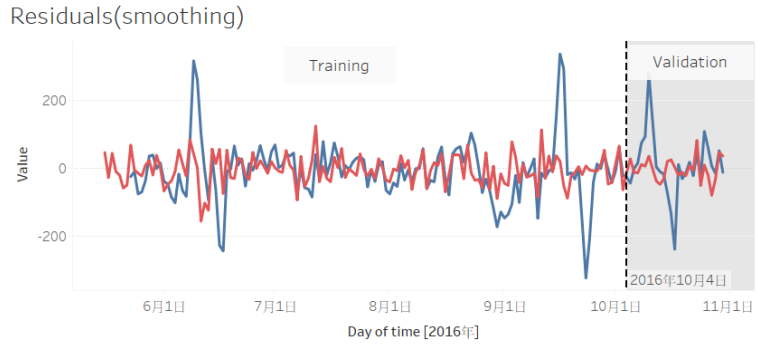




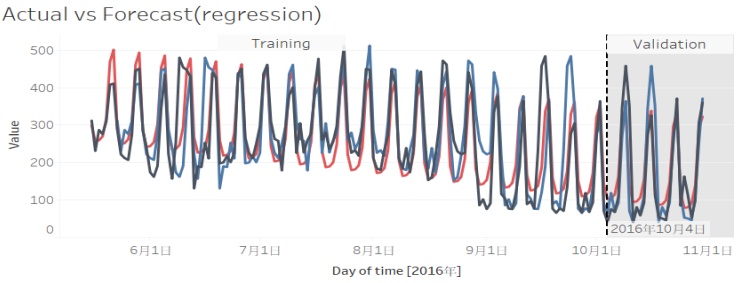
2. Smoothing

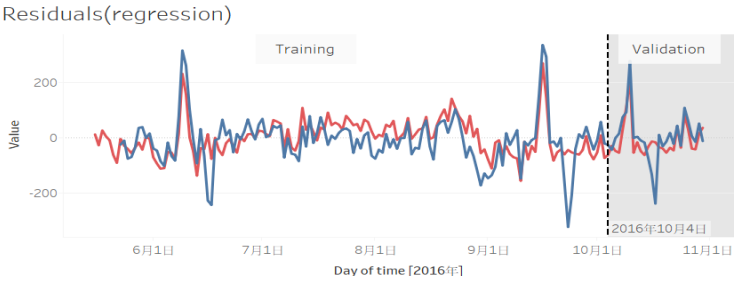


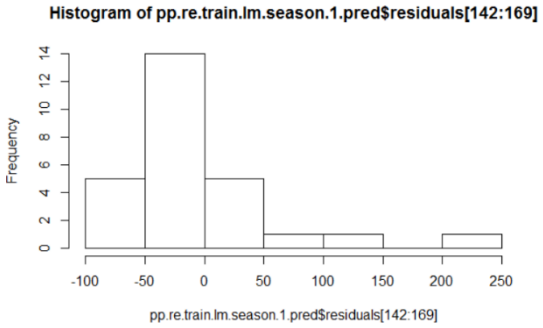




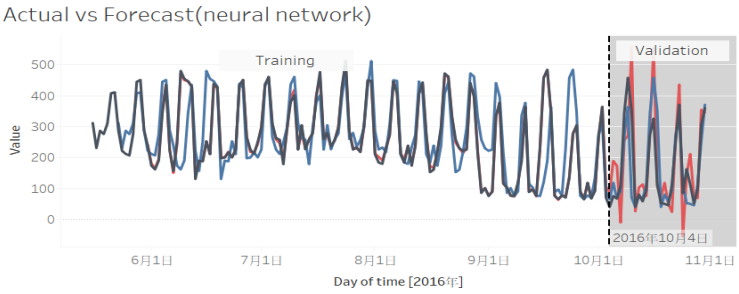
3. Regression

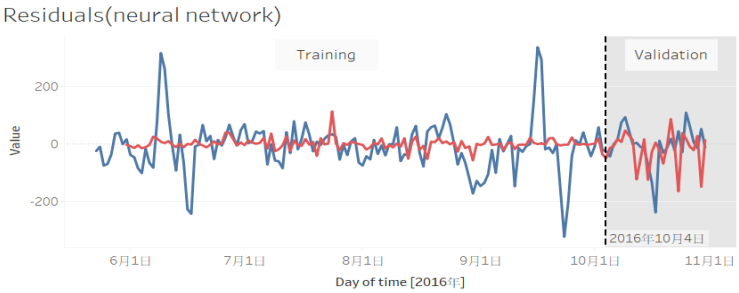


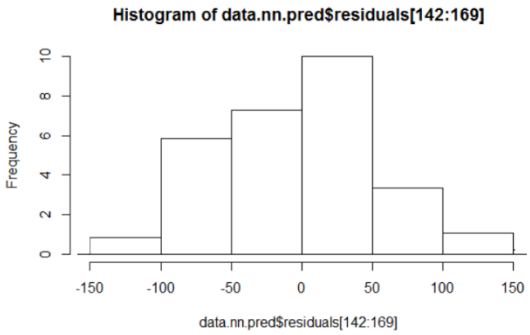




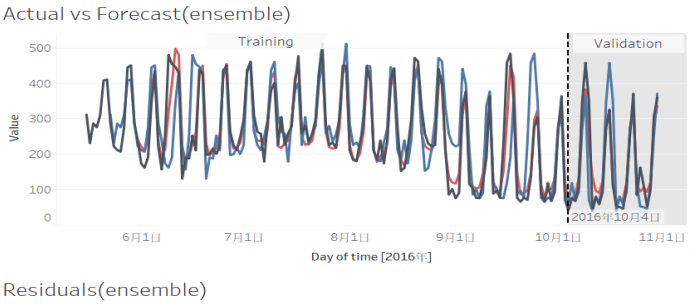
4. Neural Network

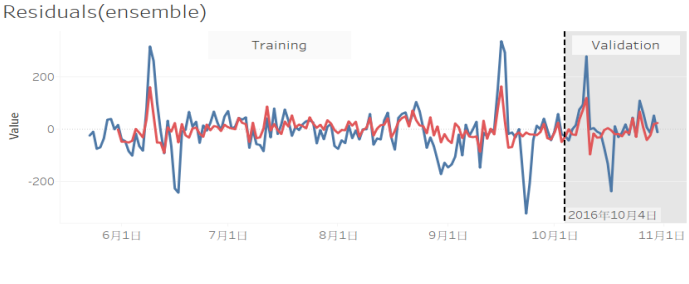


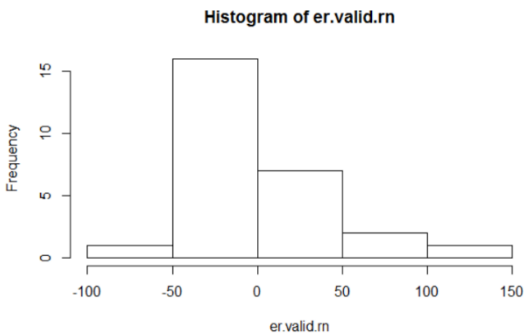




5. Ensemble (seasonal naïve + regression +neural network)

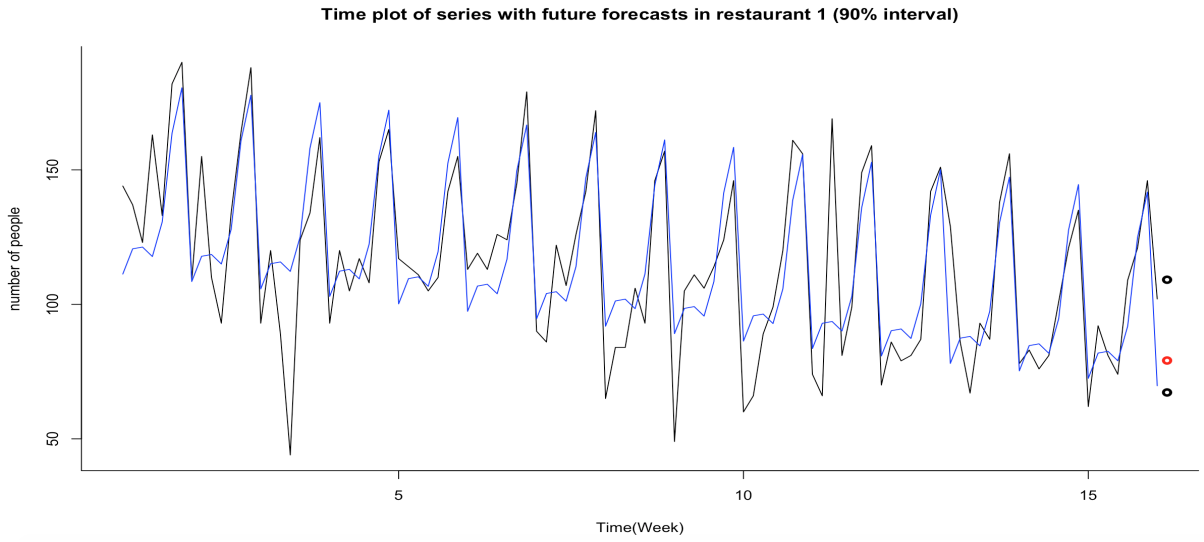


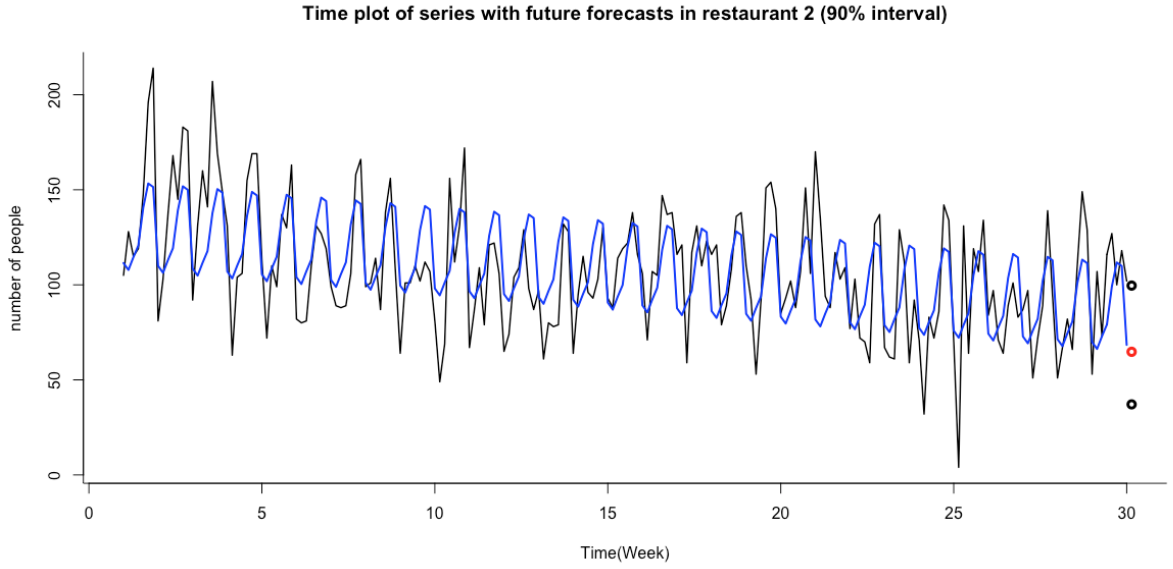


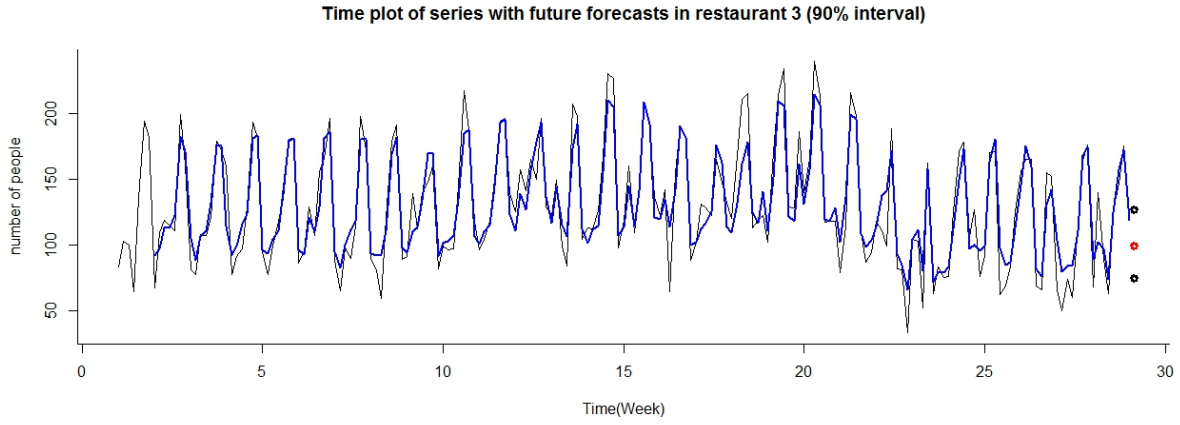


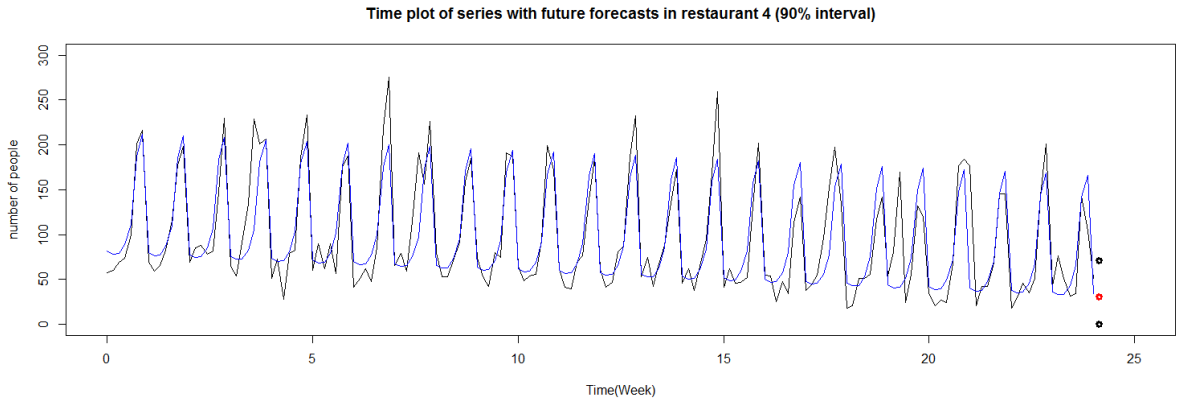
4. Time plot of series with future forecasts for each of the 5 series

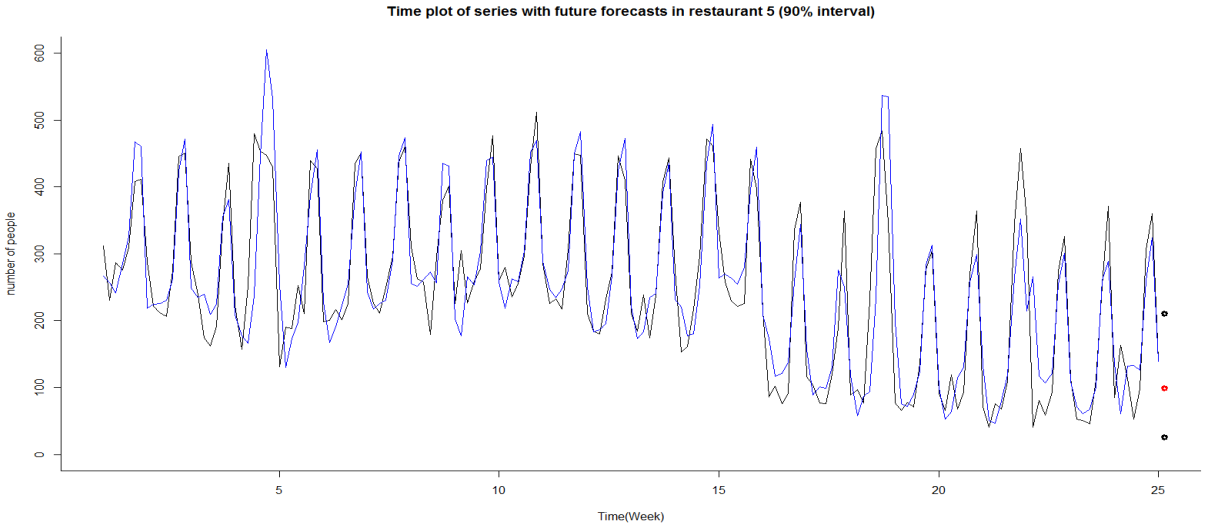
Note: Red dot means point forecast on November 1, 2016, and two black dots represent forecast uncertainty (90% confidence interval).








Time plot of series with future forecasts in restaurant 4 (90% interval)
number of people
0
50
100
150
200
250
300
0
5
10
15
20
25
Time(Week)


Time plot of series with future forecasts in restaurant 5 (90% interval)
number of people
0
100
200
300
400
500
600
5
10
15
20
25
Time(Week)