Forecasting Daily Number of User Problem Reports of Junyi Academy for Efficient Staff Allocation

均一
教育平台

Business Analytics Using Forecasting
Group 7

Chia Li Chien
Sherry Wu
Elisa Wang
Emily Wu

Executive Summary

Introduction of Our Client: Junyi Academy In this project, our client is Junyi Academy, a platform offering online learning resources for all ages. It provides practice exercises, instructional videos, and a personalized learning dashboard that empower learners to study at their own pace in and outside of the classroom. With high utilization ratio of the practice exercises, Junyi receives problems reported by users, which are called “user problem reports”. All the reports will be checked and then be distributed to the responsible team by operation team.

Business Problem The main business goal for this project is to help the manager of Junyi Academy operation team better allocate the staffs and their work loading. Since there is no full-time staff dealing with user problem reports, if daily reports are over **23**, which is the average number of reports solved, it is very likely that reports cannot be solved on that day. Our project is going to forecast daily number of user problem reports of the next week. With the forecasts, the manager of operation team can come up with proper actions to handle the days with plenty reports.

Data Description We got data from Junyi Academy, including daily number of user problem reports, daily number of active users and daily number of new registered users. The time period of the series are from Aug. 29th, 2016 to Nov. 13th, 2016, totally 77 records in each series. We also marked out the “student school day” as a series, which meant whether students had to go to school or not, and also “outlier”, which meant the usual number of the user problem reports.

Forecasting Solution Before forecasting, we applied data visualization technique to detect data patterns. Then, we chose and try several forecasting methods appropriate for our data, including seasonal naive, moving average, Holt-Winter’s smoothing (ANA), linear regression model and neural networks. Finally, considering both our goal and prediction accuracy, we decided to apply neural networks and also do roll-forward forecasting to get the prediction interval.

Recommendations According to the results, we suggested Junyi Academy could just run our predictive model on every Friday to forecast daily number of user problem reports of the next week. Thus, the manager of operation team can, based on the forecasts and prediction interval, decide whether to allocate extra staff to deal with reports or to arrange people to check the content of the questions before released which is the main cause for user problem reports.

Detailed Report

Problem Description

In this project, our client is Junyi Academy, a platform offering online learning resources for all ages. It provides practice exercises, instructional videos, and a personalized learning dashboard that empower learners to study at their own pace in and outside of the classroom. With high utilization ratio of the practice exercises, Junyi receives problems reported by users, which are called “user problem reports”. All the reports will be checked and then be distributed to the responsible team by operation team.

- **Business Goal**

  To help the manager better allocate the staffs, and even take action to prevent from receiving overloaded number of the problem reports, we decided to forecast the number of user problem reports.
  Our client are the manager of the Operation Team in Junyi, who handles the problem reports distribution. The stakeholders are Junyi employees those in charge of the problems, and users (include students, teachers), since the more reports Junyi got, the longer time they have to wait for reply.

- **Forecasting Goal**

  We attempted to forecast the daily number of user problem reports in the coming week (a-week-ahead forecasts). To know if the number is overloaded, we set the benchmark as 23 which is the average number of the reports could be resolved per day. This is a forward-looking goal, and the forecast horizon is set to be from 1 to 7((a-week-ahead forecasts). The result will be the forecasting numbers with confidence interval, thus provide the manager a more reliable number to do the staffs allocation.

Data Description

The data collected from August 29th, 2016 to November 13th, 2016 contains 77 records. In our five series of data, three series from Junyi are “Received”, “Registration” and “Active”, and the rest two binary series that we created are “School-day” and “Outlier”.

The following are the meaning of data, the sample of 10 rows per series and the time plot :

- Received: the daily number of problem report.
- Registration: the daily number of new registration users on Junyi Academy.
- Active: the daily number of active users, who finish at least one exercise at that day.
- School-day: a binary series which records whether students go to school or not.
  (1: going to school, 0: no school)
- Outlier: a binary series which records special events. (1: special event, 0: no special)

Date	Weekday	Received	Registration	Active	School-day	Outlier
2016/8/29	星期一	11	2075	4274	1	0
2016/8/30	星期二	27	1967	5192	1	0
2016/8/31	星期三	17	1303	5402	1	0
2016/9/1	星期四	24	1751	5972	1	0
2016/9/2	星期五	27	1274	4664	1	0
2016/9/3	星期六	19	756	3902	0	0
2016/9/4	星期日	25	756	4513	0	0
2016/9/5	星期一	11	1370	5814	1	0
2016/9/6	星期二	21	1439	5987	1	0
2016/9/7	星期三	22	1837	6083	1	0

Figure 1. Sample of a 10 rows per raw data series

Figure 2. Time plot for "Received", "Registration" and "Active"

Data Preparation Details

Before forecasting, we did data preprocessing to transform the raw data into an understandable format. The following are our steps:

- First, we created seasonal dummies for the series of "season" and denotes the season for each observation. We assigned value 1 to Monday, 2 to Tuesday,..., 7 to Sunday.
- Second, we added a binary series of external data "School-day" to record whether students go to school or not. (1: the day going to school; 0: the day not going to school)
- Third, we took lag-7 version of the series for "Registration" and "Active", because we can't do forecast by using future data.
- Last but not the least, we created a binary series for the outlier. From the time plot of "Received", we found that there was a high peak on September 28th. So we drew a boxplot to check if it was an outlier. We also found that on September 27th and 28th were typhoon day-off. Because typhoon day off is unpredictable, we set value 1 to mark the unexpected situation and vice versa.

Forecasting Solution

Based on our data, which have seasonality but no trend, we consider the following methods:

Method	***Detailed Settings***
Seasonal Naive	As a benchmark.
Trailing Moving Average	Window width = 7
Holt-Winter's exponential smoothing model	"ANA" model
Multiple Linear Regression model	Include 4 external series as predictors: seasonal dummies, lag-7 of active user, school day, and outlier.
Neural Network	Using seasonal dummies and all external series as predictors. (Both try automatic and manual neural networks in XLMiner.) The model with one hidden layer and 25 neurons (the default size) performs better than the model chosen automatically

Table 1. Method and detailed setting

The time plots of the actual vs. forecast value and the comparison of RMSE and MAPE for each method are in the appendix.

According to our goal, we selected NN model from other methods for two reasons:(1) The NN model got better performance of RMSE & MAPE, though it seems a little overfit. (2) However, while looking the time plot, its ability of capturing peaks is better than all the other methods we've tried. We think the NN model would more match our goal, which is to forecast the sudden occurrence of lots problem reports

Results

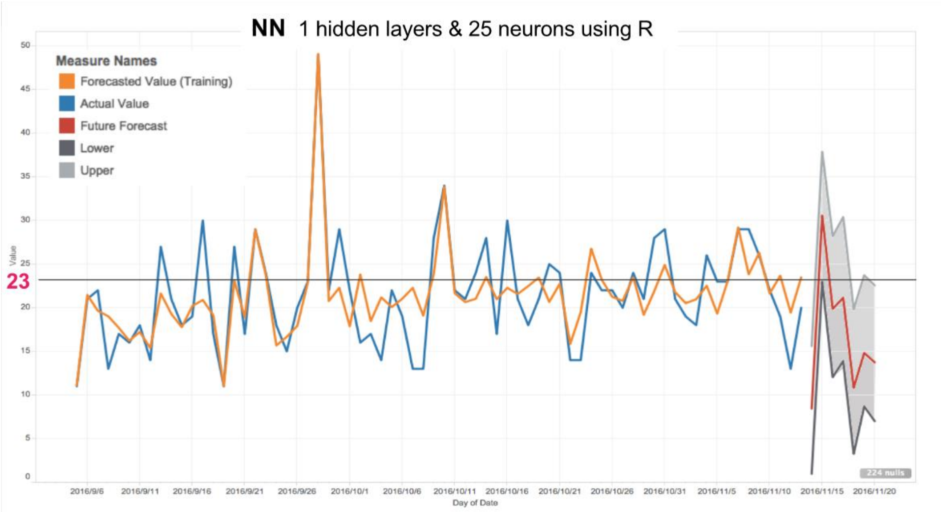

Figure 3. Time plot of actual value and the forecast by neural network

We finally use the neural network model with one hidden layer and 25 neurons to generate future forecasts and show on the graph above. The upper and lower bound of the prediction interval was calculated by the 95 percentile and 5 percentile of the forecast error in the validation period. The prediction interval can be a reference for the manager to make decision.

Conclusion

Through the whole forecasting project, we came out several limitations:

1. With the plot of the forecasts, we know that we cannot forecast the exact number of peak values, but we can capture where the peaks are.
2. After a short talk with the manager in Junyi Academy, we know that there will be lower cost for over-forecast than under-forecast. However, our model tends to be under-forecast.

According to the limitations, we suggest Junyi could just run our predictive model on every Friday to forecast the next week's report number. Thus, our forecasting can help them:

1. Decide whether to allocate extra staff to deal with reports
2. Arrange people to check the content of the questions before they are released.

Appendix

Date	Weekday	season	Received	lag7_R	lag7_A	School-day	Outlier
2016/9/5	星期一	1	11	2075	4274	1	0
2016/9/6	星期二	2	21	1967	5192	1	0
2016/9/7	星期三	3	22	1303	5402	1	0
2016/9/8	星期四	4	13	1751	5972	1	0
2016/9/9	星期五	5	17	1274	4664	1	0
2016/9/10	星期六	6	16	756	3902	1	0
2016/9/11	星期日	7	18	756	4513	0	0
2016/9/12	星期一	1	14	1370	5814	1	0
2016/9/13	星期二	2	27	1439	5987	1	0
2016/9/14	星期三	3	21	1837	6083	1	0

Figure 4. Sample of a 10 rows per raw data series after data preprocessing

	Training		Validation	
	RMSE	MAPE	RMSE	MAPE
Seasonal Naive [Benchmark]	8.8784	30.3077	6.4031	28.7795
SES	9.4985	33.1938	7.3374	28.3997
Trailing MA	10.4972	35.7355	9.1741	31.5272
Holt-Winter's (ANA)	7.0517	24.3907	6.2843	25.4447
Regression Model with external data	5.4104	21.8819	5.3927	21.5305
Neural Network (5-25-1)	3.4786	13.3547	5.1661	20.9879

Table 2. Performance of seasonal naïve, SES, trailing MA, Holt-Winter's (ANA), regression model with external data, neural network

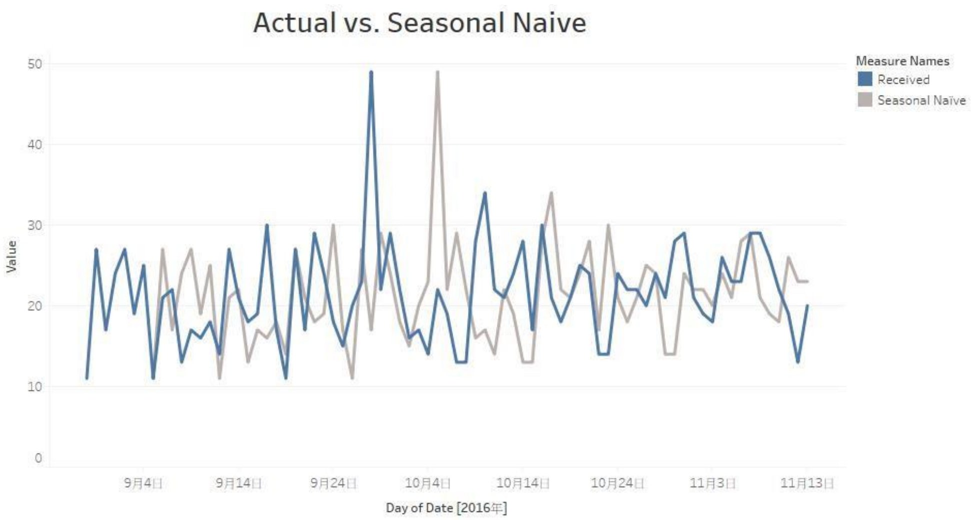

Figure 5. Time plot of actual value and the forecast by seasonal naïve (benchmark)

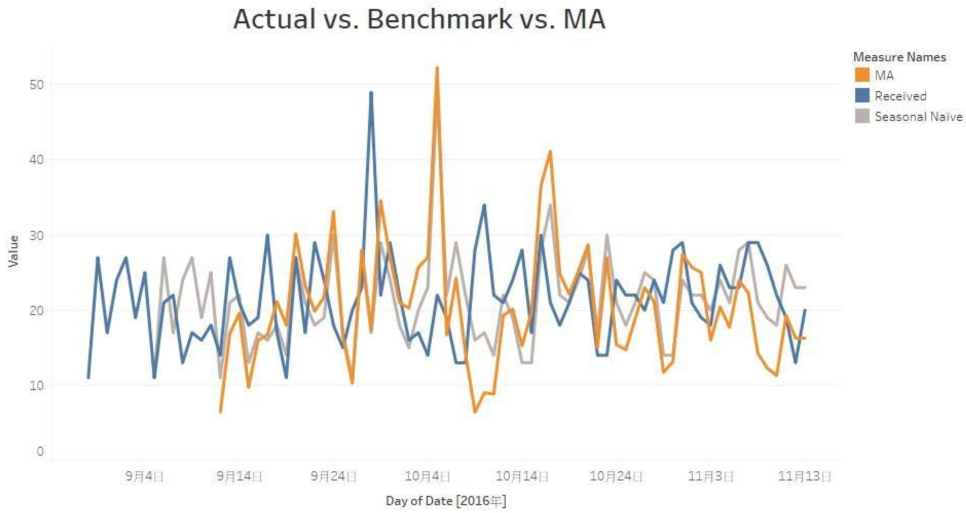

Figure 6. Time plot of actual value, benchmark and the forecast by moving average

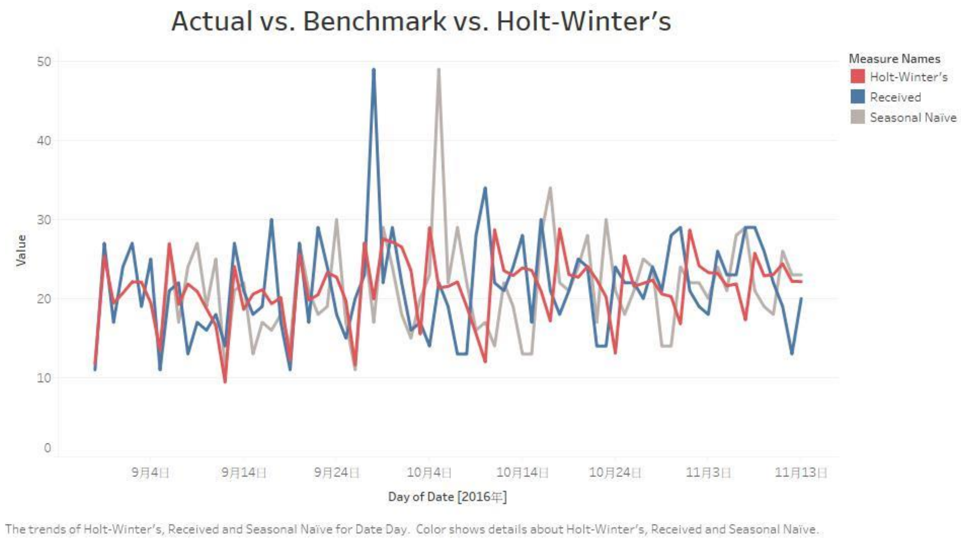

Figure 7. Time plot of actual value, benchmark and the forecast by Holt-Winter's

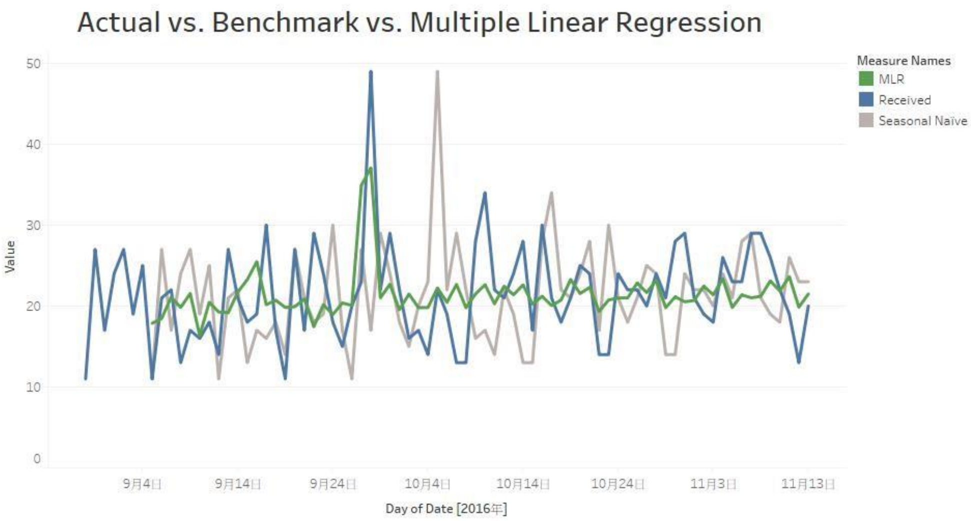

Figure 8. Time plot of actual value, benchmark and the forecast by multiple linear regression

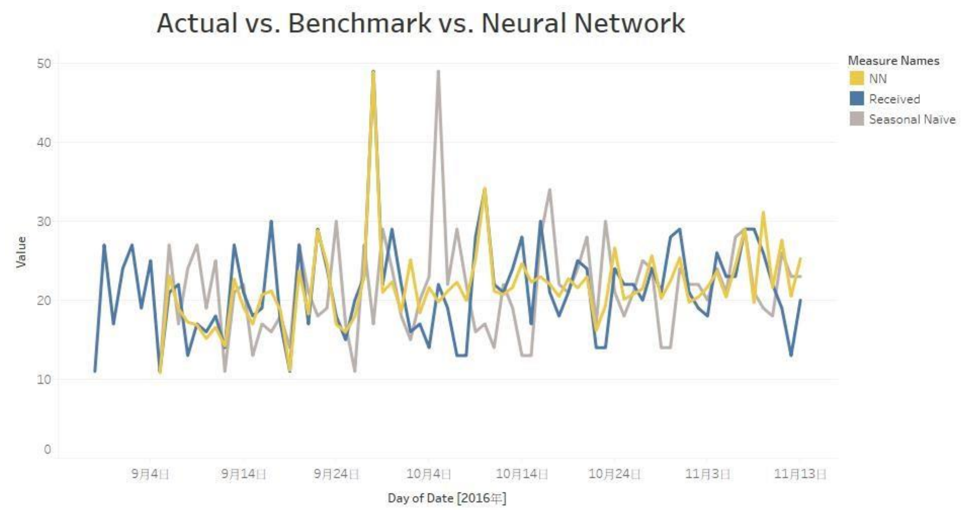

Figure 9. Time plot of actual value, benchmark and the forecast by neural network