個別部門２ 地域部門

Aakansh Gupta | 01-25-2015

A. Problem Statement:

The problem was to predict the daily number of tourists for 14 different cities of Japan from 2015-06-01 to 2015-11-31 (183 days). The number of tourists from 2014-06-01 to 2015-05-31 (365 days) was given as the training data set.

For convenience, we refer to different cities by following names:

北海道函館市	C1	静岡県熱海市	C8
宮城県仙台市	C2	三重県伊勢市	C9
東京都中央区	C3	京都府京都市	C10
神奈川県箱根町	C4	島根県出雲市	C11
神奈川県湯河原町	C5	広島県広島市	C12
富山県富山市	C6	長崎県長崎市	C13
石川県金沢市	C7	沖縄県石垣市	C14

This report focusses on 2 specific cities: C6, C7 (個別部門２) of the competition. However, because our prediction model of these 2 cities is dependent on prediction models of other cities as well, we summarize the general modelling method we used for every city. Then towards the end we show specifically the models of these 2 cities.

B. Summary of Data Sets Used:

1. Contest Data Provided:

S.No.	Data Set Name	Source
D1	SNS-Keywords Data	株式会社ホットリンク
D2	SNS-Location Data of 14 cities	株式会社ナイトレイ

D3	Exchange Rates Data	http://fx.sauder.ubc.ca/data.html
D4	Sensor Data of 49 stations	株式会社NTTドコモ
D5	Weather Data of 36 sites	株式会社NTTドコモ

2. Other Open Data Used:

S.No.	Data Set Name	Source
OP1	Geo-Coordinates of Weather Sites	www.latlong.net
OP2	Geo-Coordinates of Sensor Stations	www.latlong.net
OP3	Geo-Coordinates of 14 target Cities	www.latlong.net
OP4	Categorization of SNS keywords into 15 categories for each city (C1, C2, …, C14, GENERAL)	Manually Created
OP5	National Holidays data (Japan, China, Hong Kong, Australia, UK, US, Malaysia, Singapore, South Korea, Taiwan, Thailand)	
OP6	School Holidays data (Japan, China, Hong Kong, Australia, UK, US, Malaysia, Singapore, South Korea, Taiwan, Thailand)	
OP7	Sakura and Momiji Season in each of 14 cities	
OP8	Weekly Google Trends Data for each of 14 cities	www.google.com/trends/explore

We found that Open Data - OP8 significantly improved the accuracy of our models and was a very important part of the model. So we describe in detail how we extracted this data in the next section.

C. Google Trends Data for Toyama, Kanazawa

Google provides **weekly data** related to “Interest over time” in a given region, related to given category and filtered by given keywords. We extracted weekly Google Trends data using such queries. Different queries were used for each city Ci. Each query consists of 4 parts: Location, Time Range, Category, Keywords.

1. **Location:** We used prefectures with maximum tourist going to Ci.
2. **Time Range:** May, 2014 - November, 2015
3. **Category:** We selected from 5 options Travel, Bus and Rail, Hotels and Accomodation, Tourist Destinations, Travel Agencies.
4. **Keywords:** Google shows “Related Searches” for a given keyword at the bottom of their page. We collected all the Related Searches for city Ci in the 5 above categories. All these keywords were then combined and grouped into 4 sets: related to accommodation, related to cuisine, related to onsen/tourist spots, related to general travel/trains. We always tried to obtain a combination of very specific keywords to avoid noise being captured in the data.

The following is a summary of Google Trends data for Toyama City and Kanazawa City.

Toyama:

Only one keyword was used for Toyama : 富山市 (because sufficient Google Trends data wasn’t available for other keywords).

Location	Time Range	Category	Keyword
Japan	May,2014 - Nov,2015	Travel	富山市
Tokyo	May,2014 - Nov,2015	Travel	富山市

Kanazawa:

We used 5 keywords for Kanazawa:

1. **K1** = 金沢

2. **K2 (related to travel)** = バス金沢東京+富山金沢バス+金沢観光バス+福井金沢バス-"金沢から福井"+金沢富山電車+金沢旅行+jr バス金沢+新幹線 金沢+東京から金沢+金沢お土産+福井から金沢+金沢温泉+金沢温泉日帰り+金沢祭り+金沢観光+金沢ツアー+金沢高速バス+金沢イベント+金沢の観光+金沢名物+金沢土産+金沢市観光+金沢日帰り

3. **K3 (related to accommodation)** = jtb 金沢+テルメ 金沢+マイステイズ金沢+加賀屋 金沢+日航ホテル+東横イン 金沢+金沢 じゃらん+金沢 ホテル+金沢宿泊+金沢旅館

4. **K4 (related to food)** = 金沢カニ+金沢名物+金沢かに+金沢グルメ+金沢朝食+金沢蟹

5. **K5 (related to food + accomodation)** = 金沢カニ+金沢名物+金沢かに+金沢グルメ+金沢朝食+金沢蟹+jtb 金沢+テルメ 金沢+マイステイズ金沢+加賀屋 金沢+日航ホテル+東横イン 金沢+金沢 じゃらん+金沢 ホテル+金沢宿泊+金沢旅館

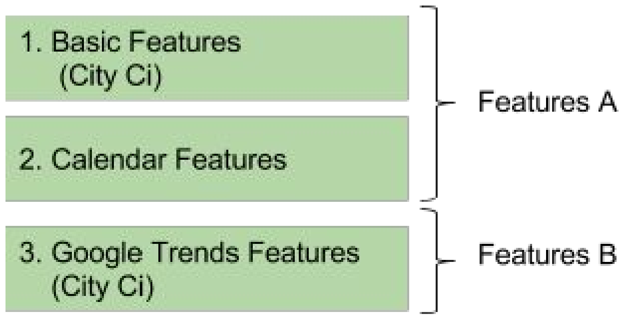
Location	Time Range	Category	Keyword

Japan	May,2014 - Nov,2015	Travel	K1
Japan	May,2014 - Nov,2015	Bus & Rail	K1
Japan	May,2014 - Nov,2015	Hotels & Accomodation	K1
Japan	May,2014 - Nov,2015	Tourist Destinations	K1
Japan	May,2014 - Nov,2015	Travel Agencies	K1
Japan, Tokyo, Kanagawa, Saitama, Osaka	May,2014 - Nov,2015	Travel	K2
Japan, Tokyo	May,2014 - Nov,2015	Bus & Rail	K2
Japan, Tokyo	May,2014 - Nov,2015	Hotels & Accommodation	K2
Japan, Tokyo, Kanagawa, Osaka	May,2014 - Nov,2015	Tourist Destinations	K2
Japan, Tokyo, Saitama	May,2014 - Nov,2015	Travel Agencies	K2
Aichi, Ishikawa	May,2014 - Nov,2015	Travel	K2
Japan, Tokyo, Kanagawa, Saitama, Osaka	May,2014 - Nov,2015	Travel	K3
Japan, Tokyo, Kanagawa, Saitama, Osaka	May,2014 - Nov,2015	Hotels & Accommodation	K3
Japan, Tokyo, Saitama	May,2014 - Nov,2015	Tourist Destinations	K3
Japan, Kanagawa, Saitama, Osaka	May,2014 - Nov,2015	Travel Agencies	K3
Aichi, Chiba,Ishikawa	May,2014 - Nov,2015	Travel	K3
Aichi, Ishikawa	May,2014 - Nov,2015	Hotels & Accommodation	K3
Japan, Tokyo, Saitama	May,2014 - Nov,2015	Travel	K4
Japan, Tokyo, Kanagawa, Saitama, Osaka	May,2014 - Nov,2015	Travel	K5
Aichi, Chiba, Ishikawa	May,2014 - Nov,2015	Travel	K5

* Note that Google Trends data is not available for prefectures with low search volume. Thus many prefectures/categories are missing in above table.

D. Explanatory Variables/Features Construction:

We constructed several features from data sets D1-D5 and open data sets OP1-OP8. For a given city Ci, the features can be broadly classified as follows.



1. **Basic Features (for city Ci):**

   a. Using open data OP4 (Categorization of SNS keywords) and D1, we picked counts of SNS keywords related to city Ci (CITYKEYWORDS_keywordXX_snsYY)

   b. sum_bbs_blog_twiiter, sun_bbs, sum_blog, sum_twitter of above.

   c. Using open data OP4 (Categorization of SNS keywords) and D1, we picked counts of SNS keywords related to category “GENERAL”. (COMMONKEYWORDS_keywordXX_snsYY)

   d. sum_bbs_blog_twiiter, sun_bbs, sum_blog, sum_twitter of above.

   e. sum_bbs_blog_twiiter, sun_bbs, sum_blog, sum_twitter of all keywords of data D1.

   f. SNS-location data (D2) counts and rowmeans.

   g. Data D3 columns.

   h. Sensor data (D4) features for 3 nearest sensor stations. (Distances were calculated using open data OP2, OP3) + statistical measures like (sensor_MaxDayTemp - sensor_MinDayTemp), RelativeHumidity_mean, RelativeHumidity_sd, TotalPrecipitation_mean, TotalPrecipitation_sd.

   i. Weather data (D5) features for 2 nearest weather sites. (categorical features were one-hot-encoded).

2. **Calendar Features:**

a. month, day of month, day of week, is_saturday, is_sunday.

b. National Holidays data + number of consecutive national holidays (OP5)

c. School Holidays data (OP6)

d. Peak Indicator { 1 if saturday, sunday, national holiday, Dec-31, Jan-1, Jan-2, Obon (08-14, 08-15, 08-16) otherwise 0 }.

e. Peak Estimate (usually peaks occur on saturdays on a weekend or on sunday on a 3-day-weekend or on monday on a 4-day weekend).

f. Trend Indicator (school holidays, Sea-Day (20 July) period, New Year period, Golden Week period, Obona period). These are days when tourism trend is usually high.

g. Moving Averages of window size 3 of Peak Indicator, Peak Estimate and Trend Indicator.

3. **Google Trends Features (for city Ci):**

a. Remove features with near zero variance (using nzv function from package caret in R).

b. **Note that to predict tourists on day d we can not use data of the week containing day d because this will be equivalent to using future data.** So, we used google trends weekly data after shifting by 1 week, 2 weeks, 3 weeks, mean of previous 2 weeks and mean of previous 3 weeks to avoid future data usage/data leakage.

c. Convert weekly features to daily features by imputing the same weekly value for each day of the week.

d. Normalize for 0 mean and 1 variance using Mean and Var of only training data set.

The features are preprocessed by removing near zero variance features and highly correlated features.

E. Overview of Models:

We created several XGBoost (Gradient Boosted Decision Tree) models for all 14 cities using different sets of features and different sets of hyperparameters. The target variable was first converted using logarithm and training was performed by using MAE (Mean Absolute Error metric).

Besides, the models were trained on either a) all 365 days (2014-06-01 to 2015-05-31) or b) with new year time (Dec, Jan, Feb) removed. This is because the number of tourists on new year rises abruptly which adds noise to predictions for the period June 2015 to November 2015.

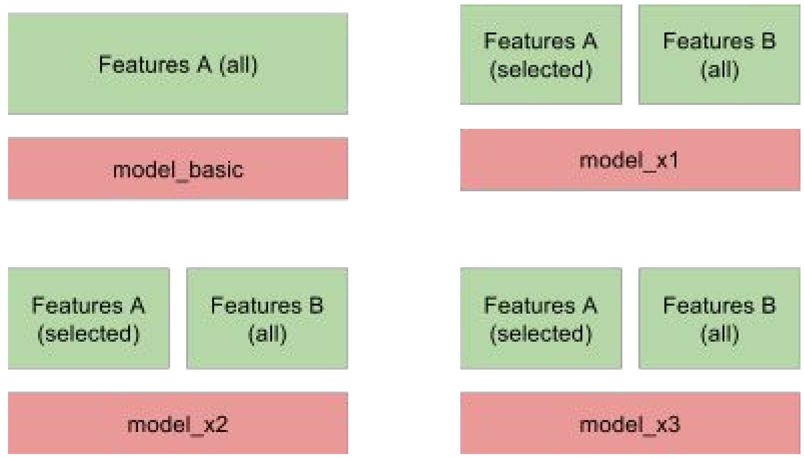
Also, the target variable can be used a) “as it is” without modifying or b) can be modified by removing the seasonal component (seasonality = 7 days) [using the stl decomposition function in forecast package in R] and then adding back the seasonality after prediction.

The various modelling possibilities are explained in the following diagram:

	A	**B**
Features	Basic Features Calendar Features (Features A)	Google Trends Features (Features B)
Target Variable	Number of Tourists	Seasonally Adjusted Number of Tourists
Training Data Size	All 365 days	All 365 days - New Year Time (Dec, Jan, Feb)

There were 4 types of XGBoost models that we created. Each of them used combination of features A and features B which were selected from “feature importance” output of several other xgboost models.

Model Name	Features Used	Target Variable	Training Data Size
model_basic	A (all)	A	A
model_x1	A (selected) + B (all)	A	A
model_x2	A (selected) + B (all)	A	B
model_x3	A (selected) + B (all)	B	B



- IMPORTANT: While selecting the models, it was important to not rely too much on leader board to avoid overfitting the data. Thus we developed an internal validation system to score the models. We trained on first 245 days and validated on next 120 days, trained on first 275 days and validated on next 90 days, then trained on first 305 days and validated on next 60 days and took the average.

The Trick Model (model_last):

We noted that tourism in some of the cities had a very high correlation. So we hypothesized that including predictions of other cities as features will improve the accuracy of the models.

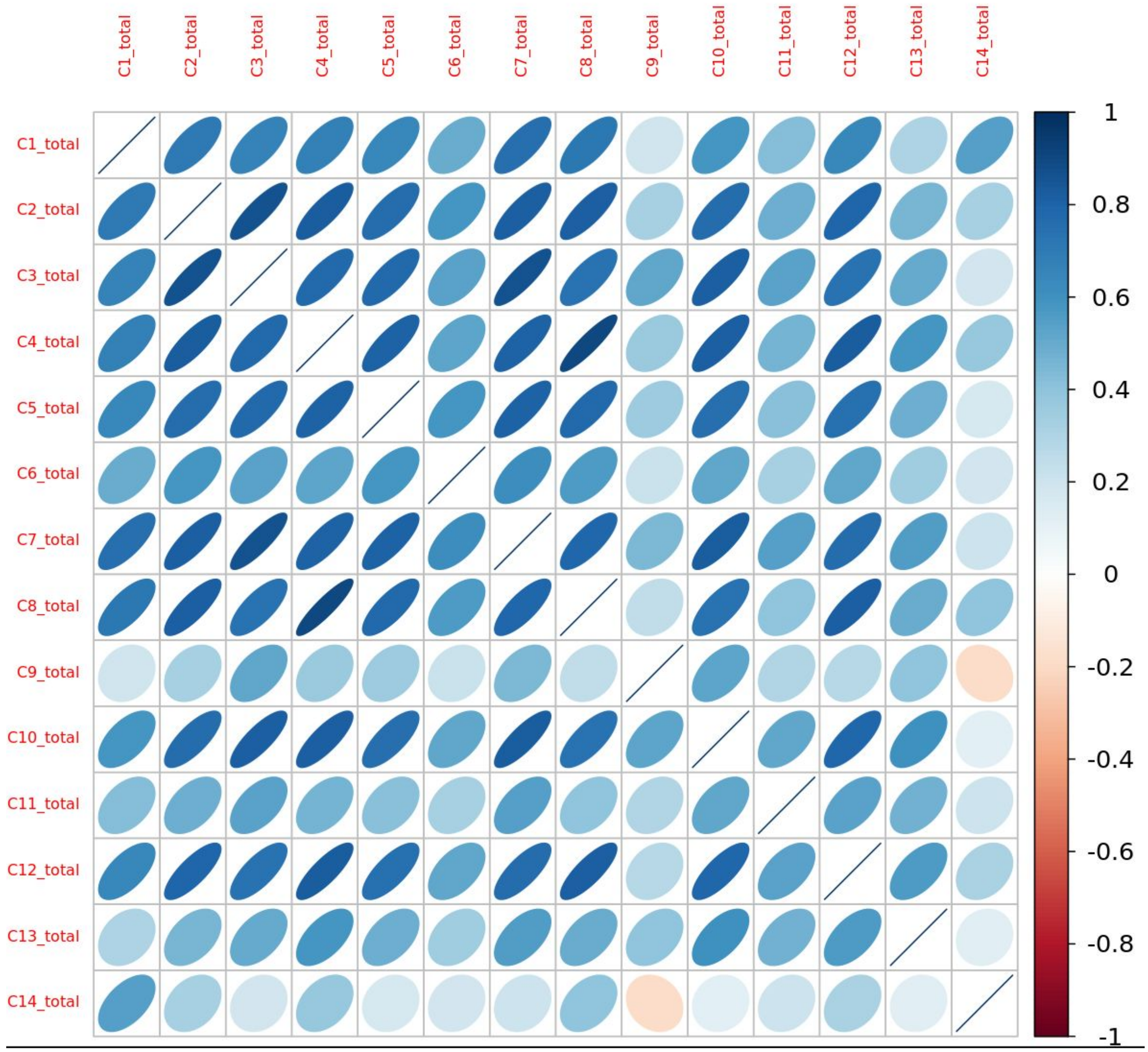
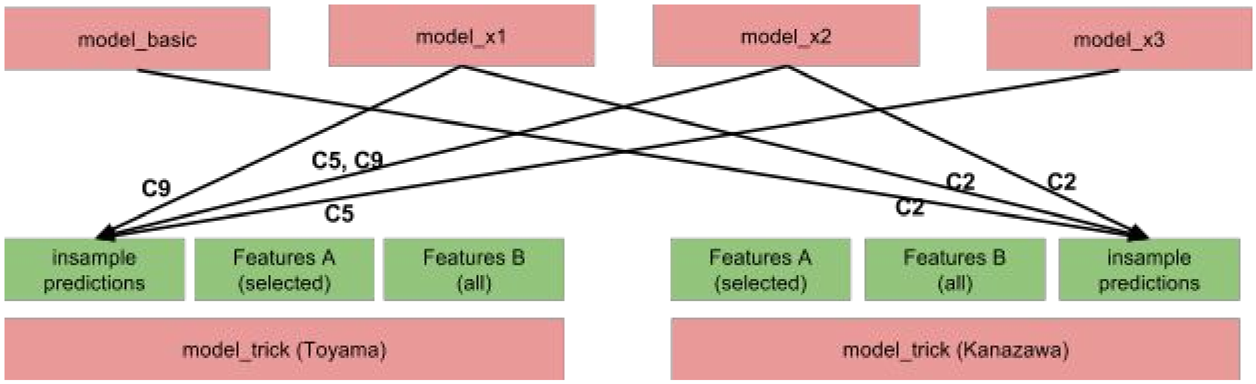


Figure 1. Correlogram of total tourists in the 14 cities for days 1:183

Precisely, for City Ci, we included the predictions of cities $Cj\ (j \neq i)$ as features in the model for city Ci. To include the predictions of other cities we had to prepare “insample predictions” as well as “out of sample predictions”. For example, we prepared “insample predictions” for days 1:183 by training on data of days 184:365 and predictions for days 184:365 by training on data of days 1:183. “Outsample” predictions were prepared by training on data of days 1:365 and predicting for days 366:548. Every time the same model parameters and features must be used.

Using these insample predictions as features, we made xgboost models to find the most important insample predictions. Finally one last single model for Toyama and Kanazawa included selected features A (Basic + Calendar Features), selected features B (Google Trends) and selected insample-prediction-features from other cities.

For cities Toyama and Kanazawa the final model looks like following:



The following table shows the internal validation scores of different models:

Internal Validation Scores of various models				
model name	Toyama City (MAE, MASE)		Kanazawa City (MAE, MASE)	
model_basic	1323.96	1.34	3111.47	1.66
model_x1	1336.37	1.35	2958.82	1.58
model_x2	1391.85	1.41	2684.64	1.43
model_x3	1340.73	1.36	**2514.48**	**1.34**
model_last	**1312.25**	**1.327**	2564.17	1.36

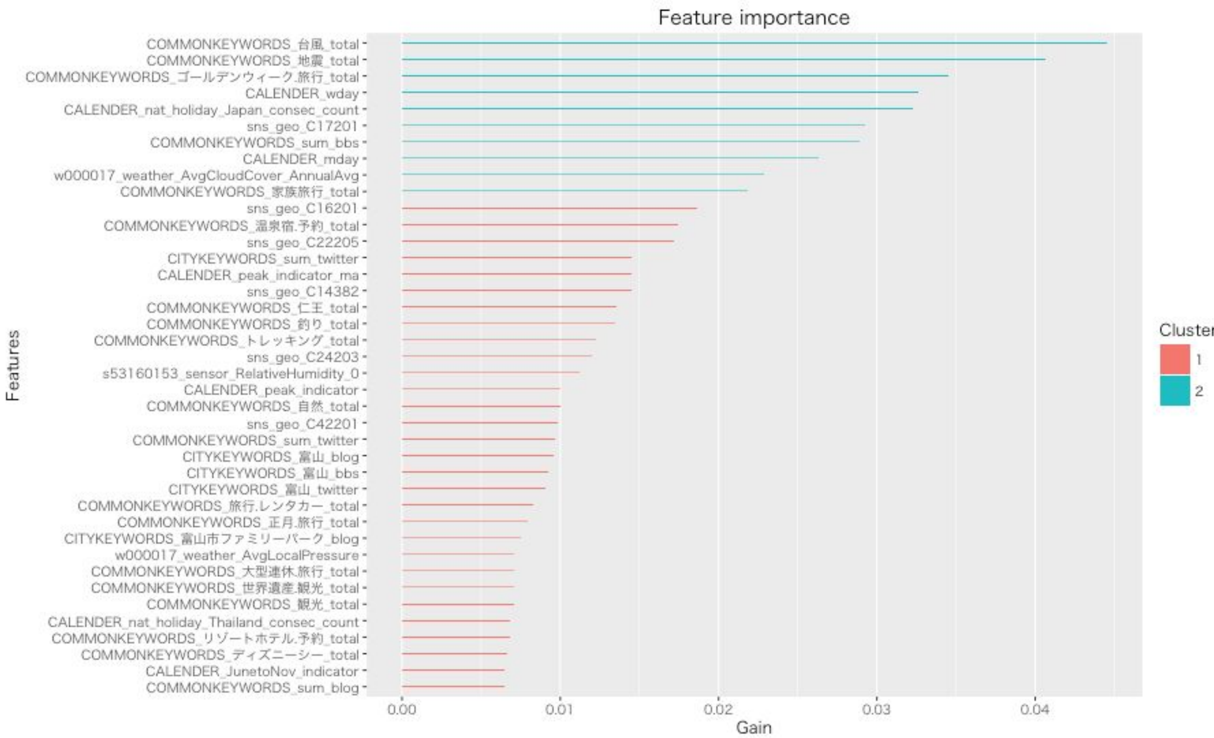
F. Final Submission

The final submission was a blended version of 2 xgboost models selected from above.

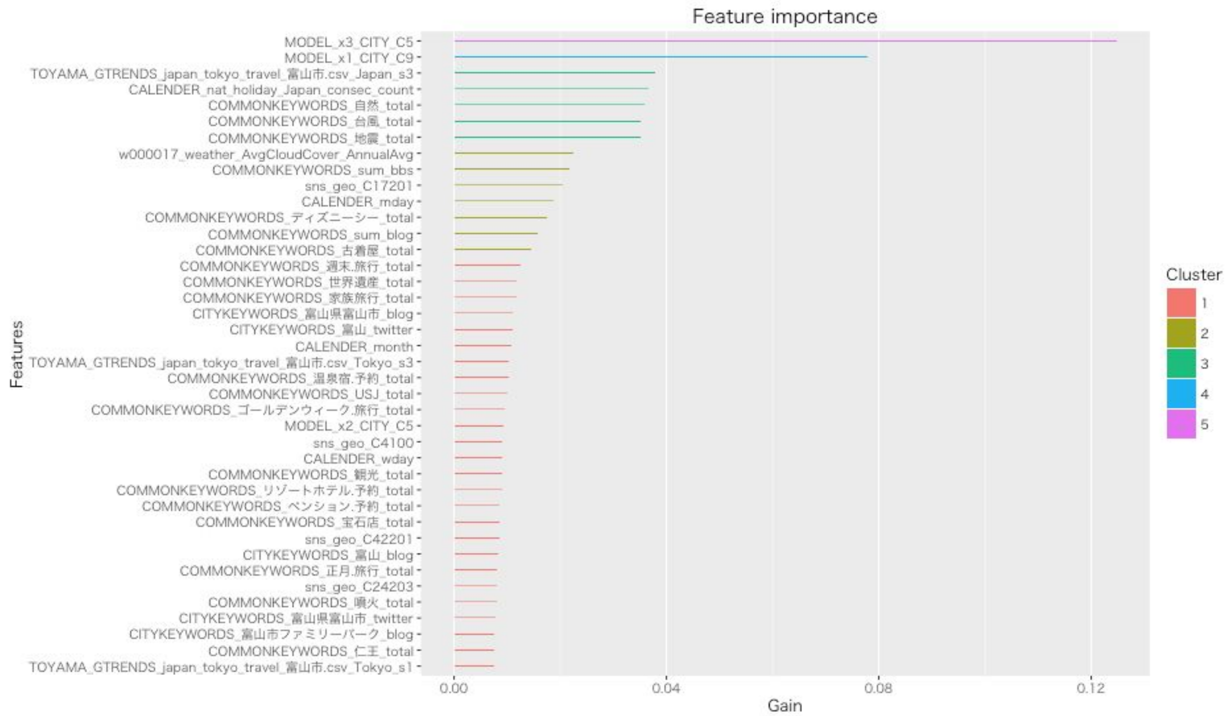
Toyama City: (model_basic + model_last) / 2
Kanazawa City: (model_x2 + model_last) / 2

G. Importance of Explanatory Variables

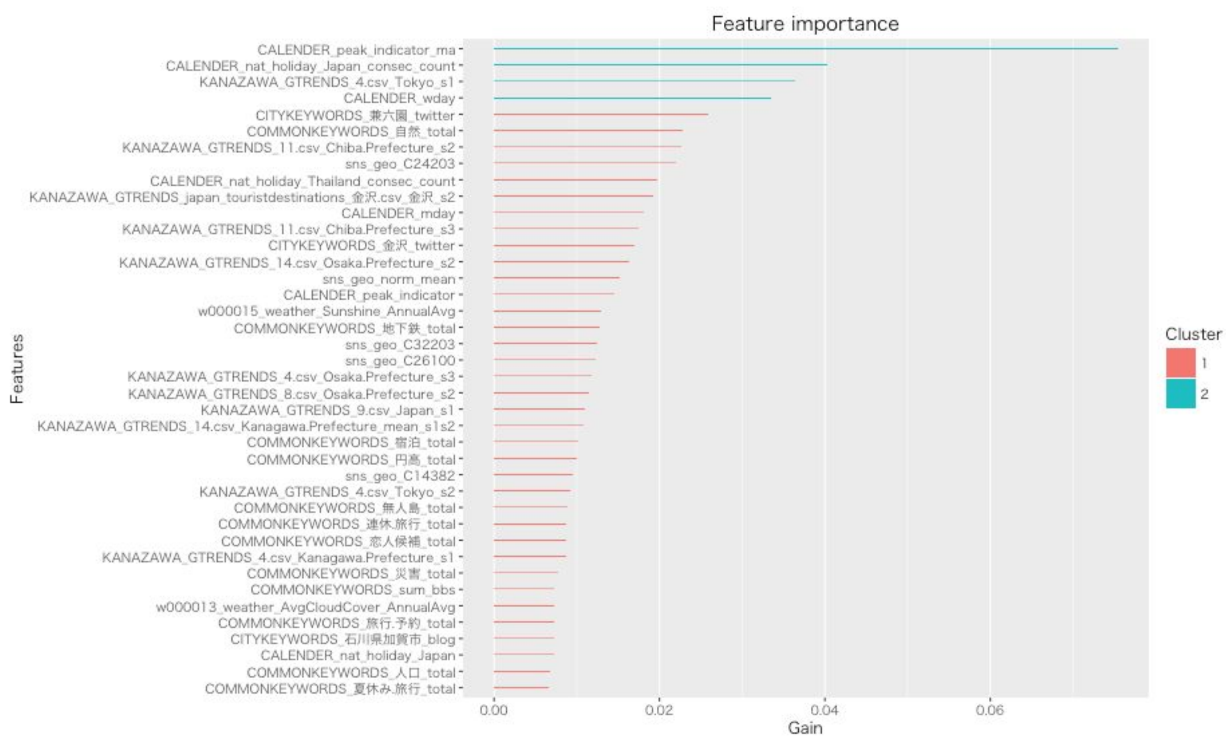
Feature Importance Plot (Top 40 features) of model_basic for Toyama:



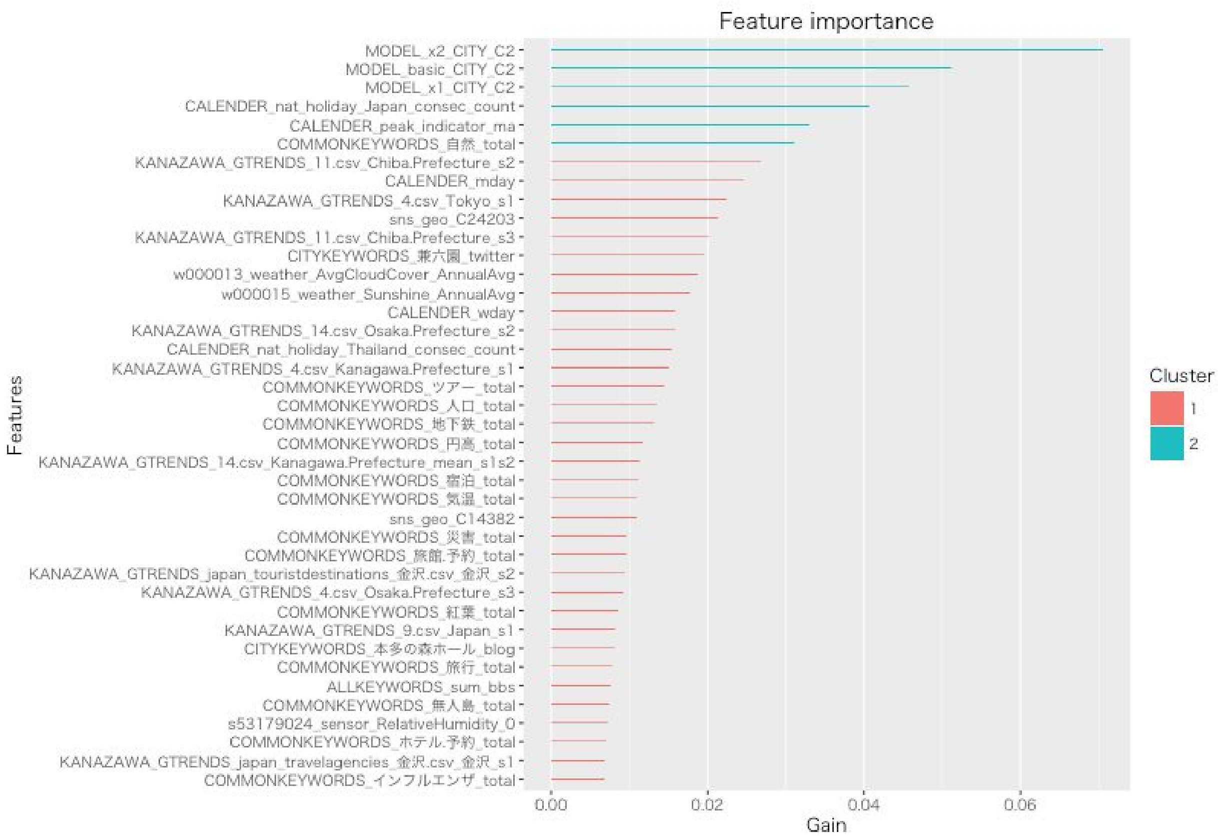
Feature Importance plot (Top 40 features) for model_last for Toyama



Feature Importance (Top 40 features) plot for model_x2 for Kanazawa



Feature importance (Top 40 features) of model_last for Kanazawa



H. Code Details

OS version, Software, modules.

- OS version: Ubuntu 14.03.3 LTS on AWS r3.8xlarge EC2 instance.
- Software: R 3.2.2 with Rstudio server.
- R packages used: data.table 1.9.6, fields 8.3-6, caret 6.0-64, reshape 0.8.5, plyr 1.8.3, forecast 6.2, xgboost 0.4-2, Metrics 0.1.1, ggplot2 2.0.0.

Xgboost Parameters.

For every model we used, eta = 0.01, subsample = 0.8, seed = 23.
Parameters are given for cities (C2, C5, C6, C7, C9).

model name	colsample	depth	rounds
model_basic	(0.6, 0.3, 0.6, 0.3, 0.3)	(5, 3, 3, 3, 5)	(662, 773, 1114, 901, 547)
model_x1	(0.6, 0.4, 0.3, 0.6, 0.5)	(5, 4, 3, 4, 4)	(1002, 941, 1192, 1044, 977)
model_x2	(0.7, 0.8, 0.3, 0.4, 0.4)	(5, 3, 3, 3, 3)	(903, 1271, 1095, 1095, 543)
model_x3	(0.4, 0.8, 0.6, 0.8, 0.5)	(7, 3, 3, 3, 3)	(846, 898, 952, 920, 503)
model_last	(NA, NA, 0.9, 0.8, NA)	(NA, NA, 3, 3, NA)	(NA, NA, 1183, 1113, NA)

Run the Code/ Model Reproducing.

1. Download the repository. (It contains code and raw data).
2. Install the required R packages as mentioned in RunMe.R
3. Inside RunMe.R modify the main_path variable to the path of the repository.
4. Source RunMe.R
5. A submission_ensemble.csv is generated containing final predictions for Toyama City (C6) and Kanazawa (C7)