個別部門２ 地域部門

Aakansh Gupta | 01-25-2015

A. 課題設定

今回の課題は、指定された日本の14都市それぞれについて以下の期間における観光客数を予測することである。
期間:2015-06-01～2015-11-31 (183日間)
モデル作成のために、2014-06-01～2015-05-31 (365日間) の観光客数がトレーニングデータとして与えられた。

以下、各都市を下表の通りに表すこととする。

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

本レポートでは、個別部門2に関連するC6とC7（金沢市、富山市）について詳述する。
ただし、上記2都市の予測モデルが他の都市の予測モデルに依存するため、最初にまず全都市についての全体的なモデリング方法について要約し、その後C6,C7の予測に特徴的な点について述べる。

B. 使用データ概要

1. 与えられたデータ

S.No.	Data Set Name	Source
D1	SNSキーワードデータ	株式会社ホットリンク
D2	SNSロケーションデータ(14都市)	株式会社ナイトレイ
D3	為替データ	http://fx.sauder.ubc.ca/data.html
D4	センサデータ(49基地)	株式会社NTTドコモ
D5	気象データ(36サイト)	株式会社NTTドコモ

2. 使用した他のオープンデータ

S.No.	Data Set Name	Source
OP1	気象データ取得サイトの地理座標	www.latlong.net
OP2	センサ情報取得基地の地理座標	www.latlong.net
OP3	目的14都市の地理座標	www.latlong.net
OP4	SNSキーワードのカテゴリ分け (C1, C2, …, C14, GENERAL)	Manually Created
OP5	国民の祝日データ (日本、中国、香港、オーストラリア、UK、US、マレーシア、シンガポール、韓国、台湾、タイ)	Google検索
OP6	学校の休日データ (日本、中国、香港、オーストラリア、UK、US、マレーシア、シンガポール、韓国、台湾、タイ)	Google検索
OP7	目的14都市それぞれの桜と紅葉の見頃データ	Google検索
OP8	目的14都市それぞれについてのGoogle Trendsデータ （週ごとのデータ）	www.google.com/trends/explore

我々が検証を行うなかで、OP8が著しく精度を上げる重要なデータであることがわかった。したがって次のパートではどのようにOP8を抽出したかを詳述する。

C. 富山市／金沢市のGoogle Trendsデータ

Google Trendsは“興味の経時変化”に関する週毎のデータを提供しており、取得データは地域、カテゴリ、キーワードで絞り込むことができる。我々は各都市（*Ci*）毎に異なるクエリを作成し、週毎のGoogle Trendsデータを抽出した。

各クエリは以下の4要素により構成される。

1. **地域:** Ci への観光客が最も多い地域、国
2. **期間:** 2014年5月 - 2015年11月
3. **カテゴリ:** ‘旅行’、’バス、電車’、 ‘ホテル、宿泊施設’ 、‘観光名所’ 、‘旅行代理店、旅行サービス’ の5つから選択
4. **キーワード:** Googleは “関連キーワード”をGoogle Trends検索結果に表示する。我々は Ci についての表示された全ての関連キーワードを上記の5カテゴリについて収集し、収集したキーワードを結合し以下の4つのグループに分けた。’宿泊に関連したキーワード’、‘食べ物に関連したキーワード’、‘温泉や観光地に関連したキーワード’、‘旅行と交通機関に関連したキーワード’の4つである。ノイズが入るのを防ぐため、キーワードの結合はより特徴的になるように行った。

以下は富山市と金沢市についてのGoogle Trendsデータのサマリである。

富山市

富山市については一つのキーワード’富山市’のみを用いた。（他のキーワードでは十分なデータ量がなかったため）

地域	**期間**	**カテゴリ**	**キーワード**
日本	2014年5月 - 2015年11月	旅行	富山市
東京	2014年5月 - 2015年11月	旅行	富山市

金沢市

金沢市については5個のキーワードを用いた。

1. **K1** = 金沢

2. **K2 (旅行と交通機関に関連したキーワード)** = 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 (宿泊施設に関連したキーワード)** = 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 (食べ物に関連したキーワード)** = 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 (食べ物と宿泊に関連したキーワード)** = 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

地域	**期間**	**カテゴリ**	**キーワード**
日本	2014年5月 - 2015年11月	旅行	K1
日本	2014年5月 - 2015年11月	バス、電車	K1
日本	2014年5月 - 2015年11月	ホテル、宿泊施設	K1
日本	2014年5月 - 2015年11月	観光名所	K1
日本	2014年5月 - 2015年11月	旅行代理店、旅行サービス	K1
日本、東京、神奈川、埼玉、大阪	2014年5月 - 2015年11月	旅行	K2
日本、東京	2014年5月 - 2015年11月	バス、電車	K2
日本、東京	2014年5月 - 2015年11月	ホテル、宿泊施設	K2
日本、東京、神奈川、大阪	2014年5月 - 2015年11月	観光名所	K2
日本、東京、埼玉	2014年5月 - 2015年11月	旅行代理店、旅行サービス	K2
愛知、石川	2014年5月 - 2015年11月	旅行	K2
日本、東京、神奈川、埼玉、大阪	2014年5月 - 2015年11月	旅行	K3
日本、東京、神奈川、埼玉、大阪	2014年5月 - 2015年11月	ホテル、宿泊施設	K3

日本、東京、埼玉	2014年5月 - 2015年11月	観光名所	K3
日本、神奈川、埼玉、大阪	2014年5月 - 2015年11月	旅行代理店、旅行サービス	K3
愛知、千葉、石川	2014年5月 - 2015年11月	旅行	K3
愛知、石川	2014年5月 - 2015年11月	ホテル、宿泊施設	K3
日本、東京、埼玉	2014年5月 - 2015年11月	旅行	K4
日本、東京、神奈川、埼玉、大阪	2014年5月 - 2015年11月	旅行	K5
愛知、千葉、石川	2014年5月 - 2015年11月	旅行	K5

* Google Trendsデータは、十分なボリュームがない抽出条件については取得できない。

D. 説明変数群の構築

我々は与えられたデータD1-D5とオープンデータOP1-OP8を用いて幾つかの説明変数群を構築した。都市 Ci に関して、説明変数群は概して以下のように分類される。

1. **基本説明変数 (都市 Ci)**

    a. OP4 (SNSキーワードのカテゴリ分け)とD1を用い、都市 Ci に関連したSNSキーワードの出現数(CITYKEYWORDS_keywordXX_snsYY)

    b. a.のキーワードについての、sum_bbs_blog_twiiter, sum_bbs, sum_blog, sum_twitter

    c. OP4 (SNSキーワードのカテゴリ分け) とD1を用い、“GENERAL” に関連したSNSキーワードの出現数 (COMMONKEYWORDS_keywordXX_snsYY)

    d. c.のキーワードについての、sum_bbs_blog_twiiter, sum_bbs, sum_blog, sum_twitter

    e. D1のキーワード全てについての、sum_bbs_blog_twiiter, sum_bbs, sum_blog, sum_twitter

    f. SNSロケーションデータ(D2) とその行平均

    g. 為替データ(D3)

    h. 都市 Ci に最も近い3つの基地のセンサデータ(D4)
    (距離はOP2, OP3を用いて計算)
    + それらの統計値(sensor_MaxDayTemp - sensor_MinDayTemp, RelativeHumidity_mean, RelativeHumidity_sd, TotalPrecipitation_mean, TotalPrecipitation_sd)

    i. 都市 Ci に最も近い2つの天候観測サイトの天候 (D5) (天候分類はone-hotエンコード済)

2. **カレンダー説明変数**

a. 月、日、曜日、土曜日か?、日曜日か?

b. 国民の祝日データと国民の祝日による連休の数 (OP5)

c. 学校の休日データ (OP6)

d. ピークインジケータ(土曜日、日曜日、国民の祝日、大晦日、元日、1月2日、お盆(08-14, 08-15, 08-16)は1、それ以外は0)

e. 推定ピーク (2連休の週末なら土曜日、3連休なら日曜日、4連休なら月曜日)

f. トレンドインジケータ (学校の休日、海の日(7/20前後)の連休、年末年始、GW、お盆休み)※これらの連休は観光動向が非常に活発である。

g. ウィンドウサイズ3の、d、e、fの移動平均

3. **Google Trends説明変数(都市 Ci)**

a. 分散が0に近い説明変数を除去 (Rのcaretパッケージのnzv関数を使用)

b. **dという日付の観光客数を予測するためにdを含む週のデータを用いることは未来のデータを使用するのと等価であり、未来のデータを使用することとデータ漏れを防ぐため、**我々はGoogle Trendsデータを以下のように時間的にシフトして用いた : 1週シフト、2週シフト、3週シフト、過去2週の平均、過去3週の平均

c. 週毎に得られた値を、日毎のデータに変換 (同じ週毎の値をその週のそれぞれの日に帰属)

d. トレーニングデータの平均と分散を用いて、平均0、分散1の値に規格化

更に相関の非常に高い説明変数を除去した。

E. モデル概要

我々は14都市それぞれに対して、異なる説明変数のセットと異なるハイパーパラメータを用い、複数のXGBoost (Gradient Boosted Decision Tree)モデルを作成した。目的変数を対数変換し、MAE (Mean Absolute Error metric)を用いてトレーニングを行った。

その上で、モデルのトレーニングを以下の2種類の期間において行った。
a) 全てのトレーニング期間(2014-06-01～2015-05-31)
b) 新年付近（12月、1月、2月）を除いた期間

これは、新年付近の観光客数が急激に上がっているため、予測にノイズを載せるからである。

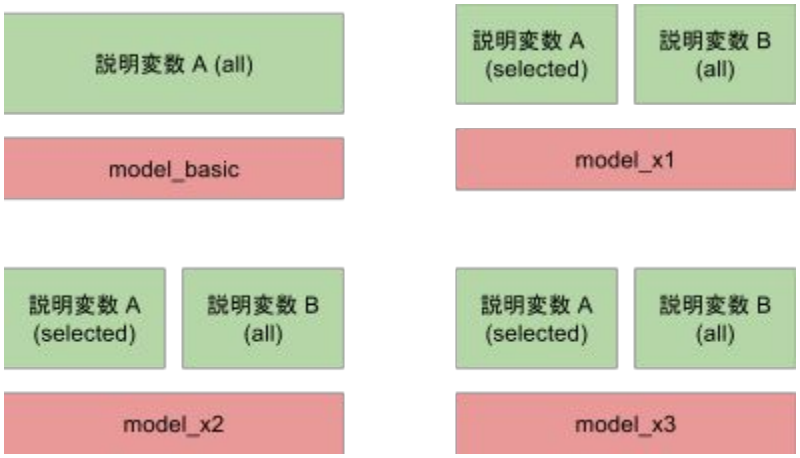
また、目的変数は以下の2種類の方法で使用した。
a) そのまま使用
b) 周期的変動要素を除去して使用（周期＝7日）[Rのforecastパッケージにあるstl decomposition関数使用] し予測実施後に周期性を追加

モデリングの選択肢は以下のマトリックスで表される。

	A	B
説明変数	基本説明変数 カレンダー説明変数 (説明変数A)	Google Trends説明変数 (説明変数B)
目的変数	観光客数	周期要素を抜いた観光客数
トレーニングデータ	365日全て	365日 - 新年期間(12月、1月、2月)

今回我々は4種類のXGBoostモデルを作成した。それぞれのモデルが、他のxgboostモデルから出力された“説明変数の寄与度” から選択された説明変数Aと説明変数Bの結合を用いている。

モデル名称	**説明変数**	**目的変数**	**トレーニングデータ**
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



- 重要:モデルの選択にあたっては、オーバーフィッティングを避けるため、リーダーボードを信用し過ぎないことが重要であった。そのため、我々は内部に評価システムを構築し、モデルの評価をおこなった。 最初の245日分のデータを用いて次の120日分の評価を行ったものと、最初の275日分のデータを用いて次の90日分の評価を行ったものと、最初の305日分のデータを用いて次の60日分の評価を行ったものの平均で評価を行った。

トリックモデル (model_last)

我々は幾つかの都市の観光客数が互いに高い相関を持つことを発見した。そのため、他の都市の観光客数を説明変数に加えることでモデルの精度が上がるのではないかと考えた。

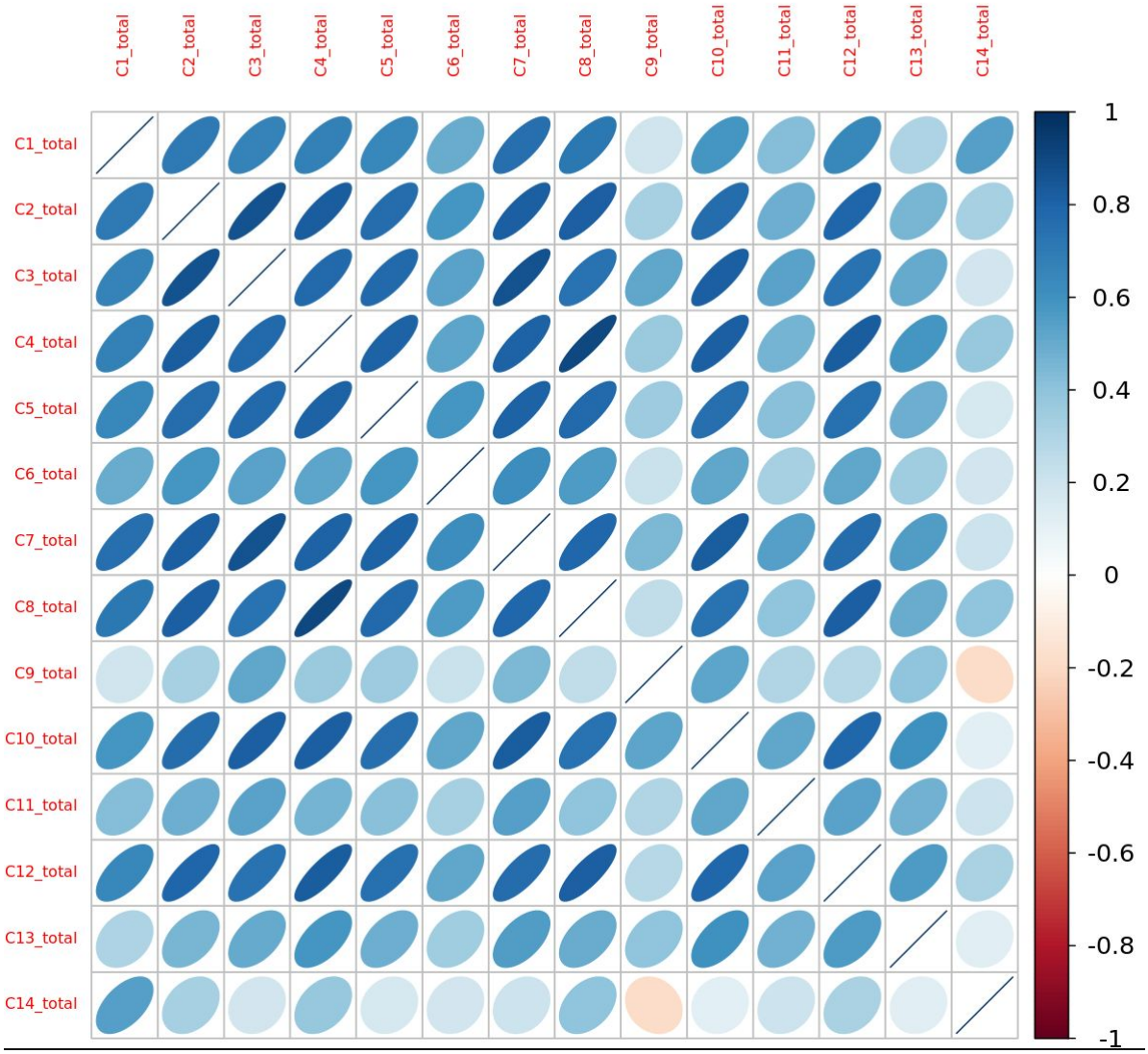
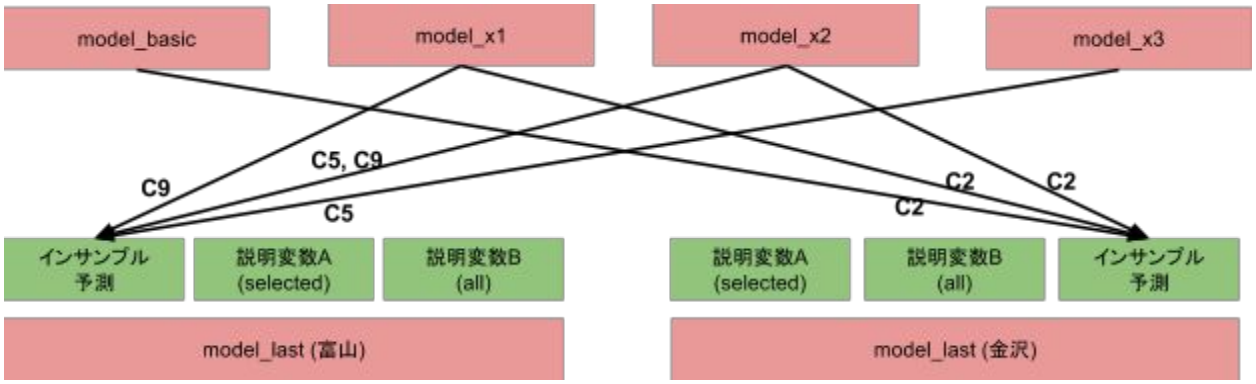


図1. 14都市間の1日目～183日目の観光客数コレオグラム

正確には、都市 Ci のモデルに都市 $Cj\ (j \neq i)$ の観光客数予測を説明変数として加えた。他の都市の予測を加える際、“インサンプル予測”と“アウトオブサンプル予測”を準備した。例えば、1日目から183日目までの“インサンプル予測”を184日目から365日目までをトレーニングデータとして行い、184日目から365日目までの予測を1日目から183日目までをトレーニングデータとして用いて行った。“アウトオブサンプル予測”は、366日目から548日目までの予測を、1日目から365日目までのデータをトレーニングデータとして予測し用意した。毎回の予測において、モデルのパラメータと説明変数は同じものを使用した。

これらのインサンプル予測を説明変数として用い、最も重要なインサンプル予測を決定するためのxgboostモデルを作成した。
最終的に、説明変数A、説明変数B、他の都市の予測を最適に含む、富山市と金沢市の予測モデルを決定した。

最終的な富山市と金沢市モデルは以下の図で表される。



下の表は、各モデルの内部評価スコアを表したものである。

内部評価スコア				
モデル名称	富山市(MAE, MASE)		金沢市 (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

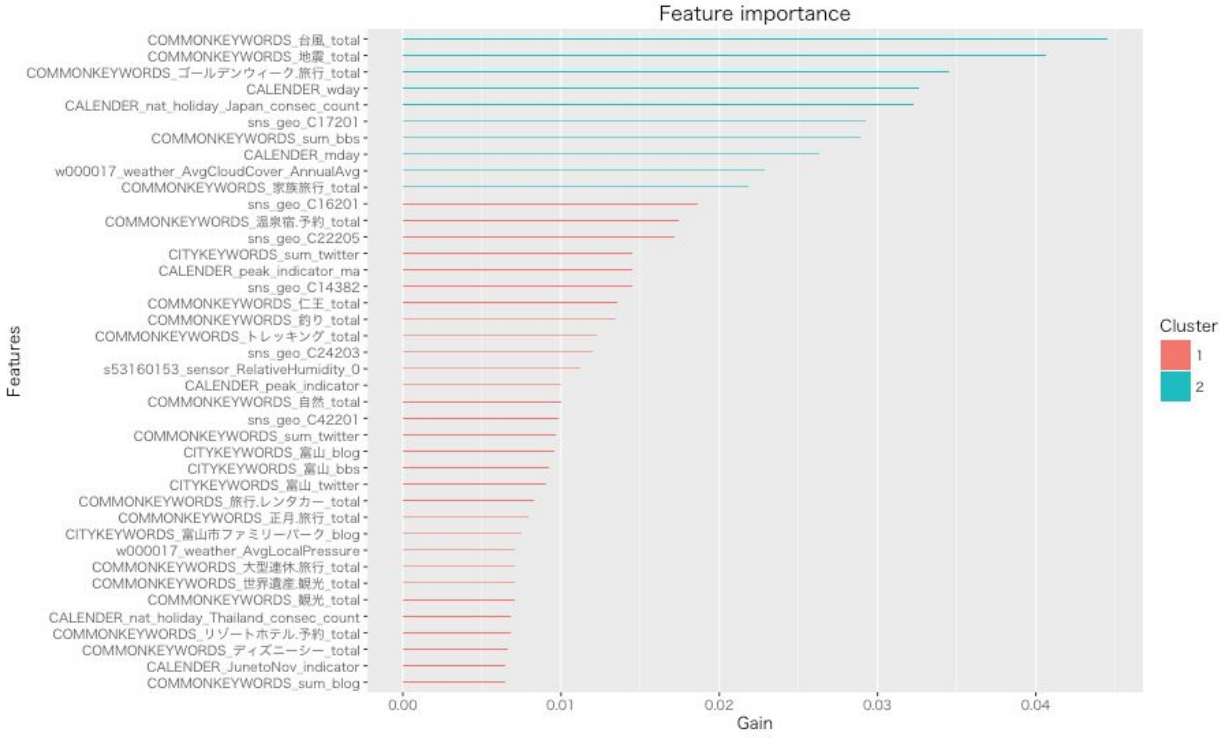
F. 最終提出モデル

最終提出モデルは、ここまで計算してきたモデルの中から2つを選択し、混成して作成した。

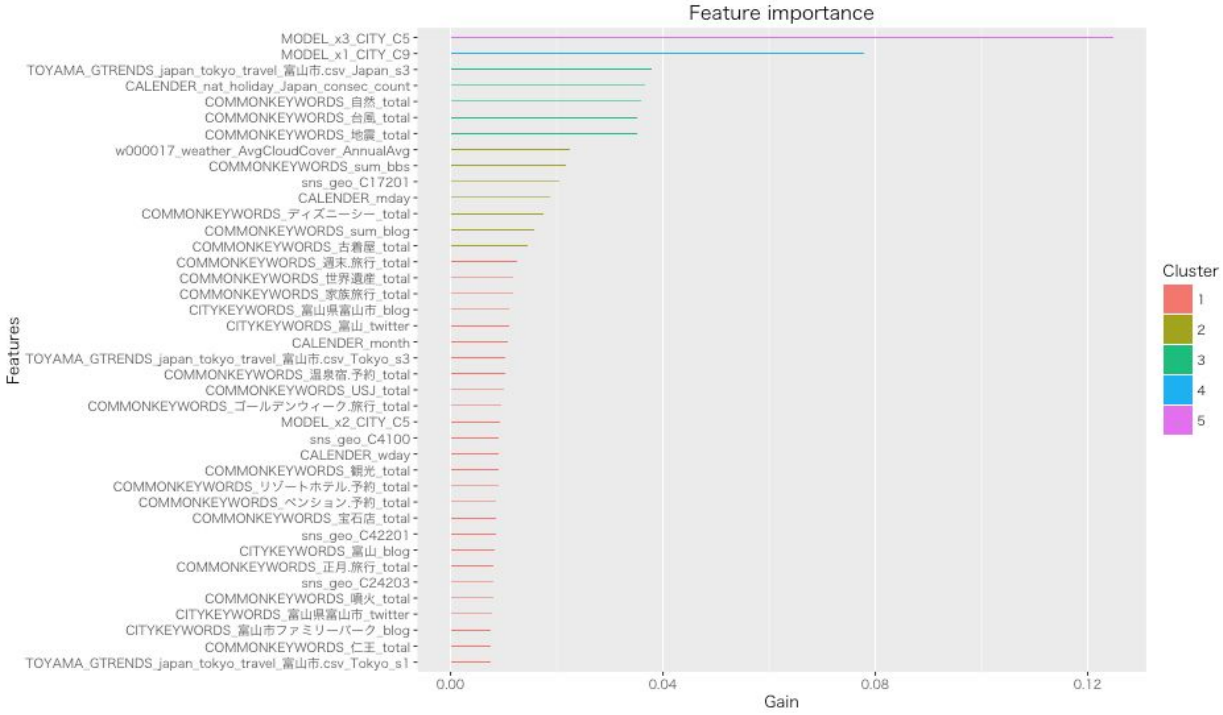
富山市: (model_basic + model_last) / 2
金沢市: (model_x2 + model_last) / 2

G. 説明変数の寄与度

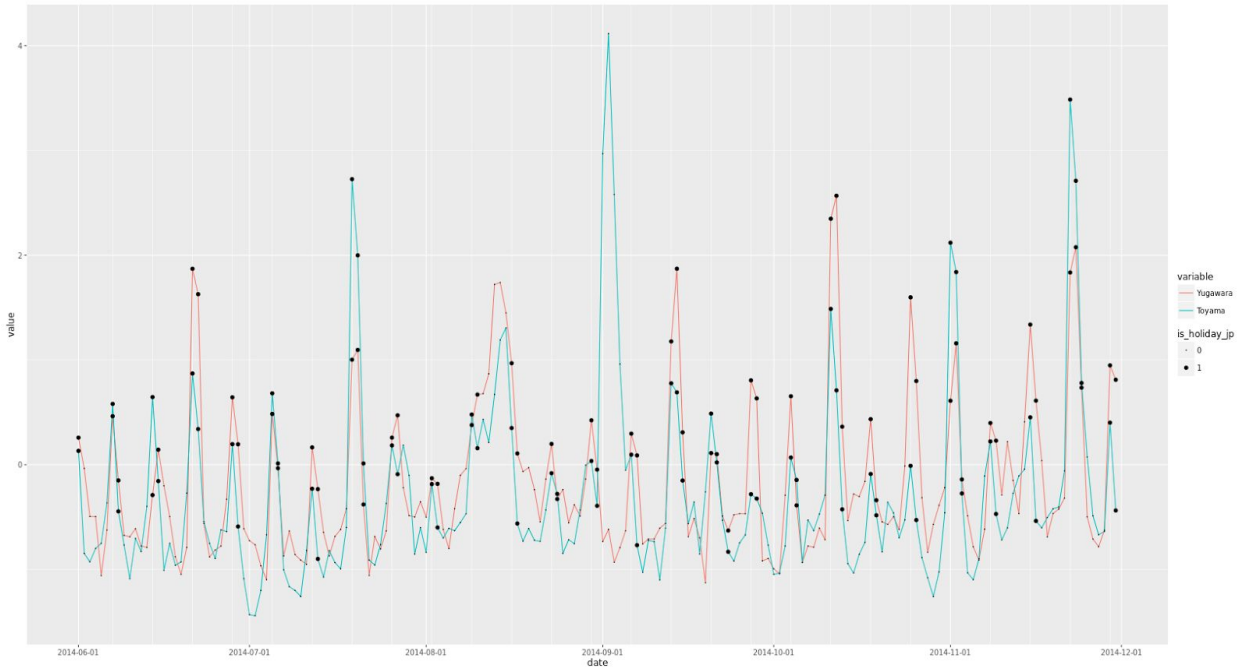
富山市のmodel_basicにおける説明変数の寄与度（トップ40）



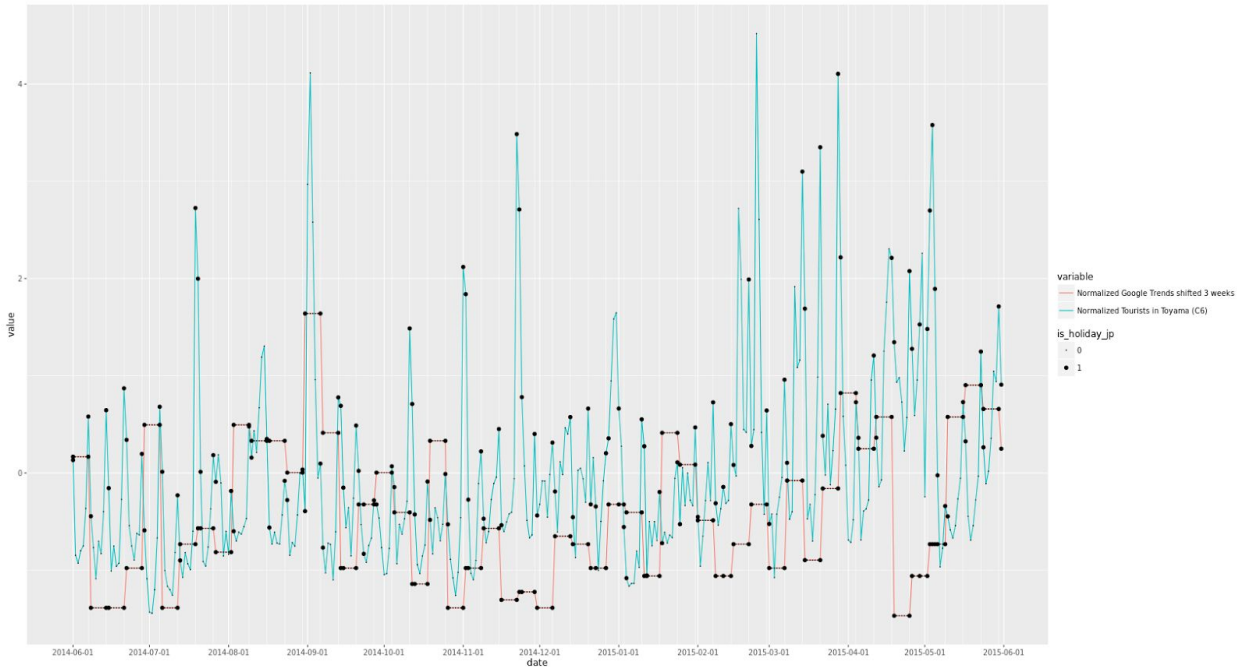
富山市のmodel_lastにおける説明変数の寄与度（トップ40）



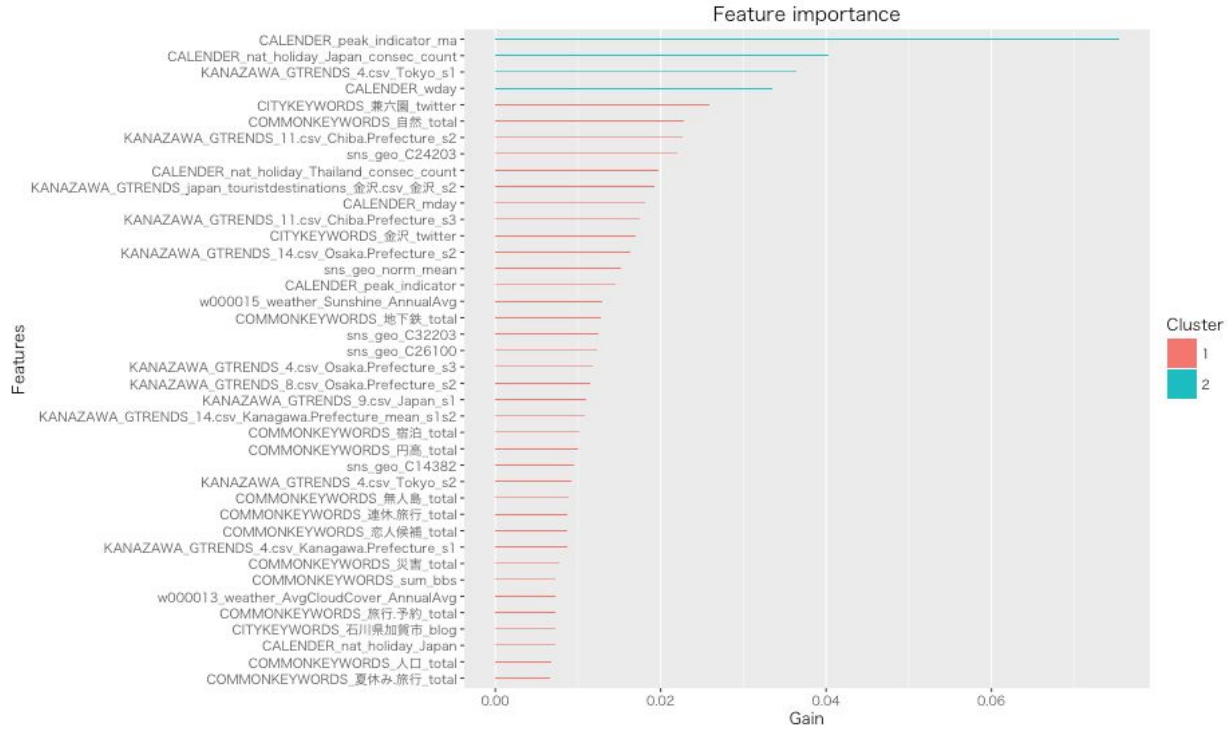
正規化した湯河原(C5) と富山市(C6) (2014年6月- 2014年11月) の観光客数傾向の相似



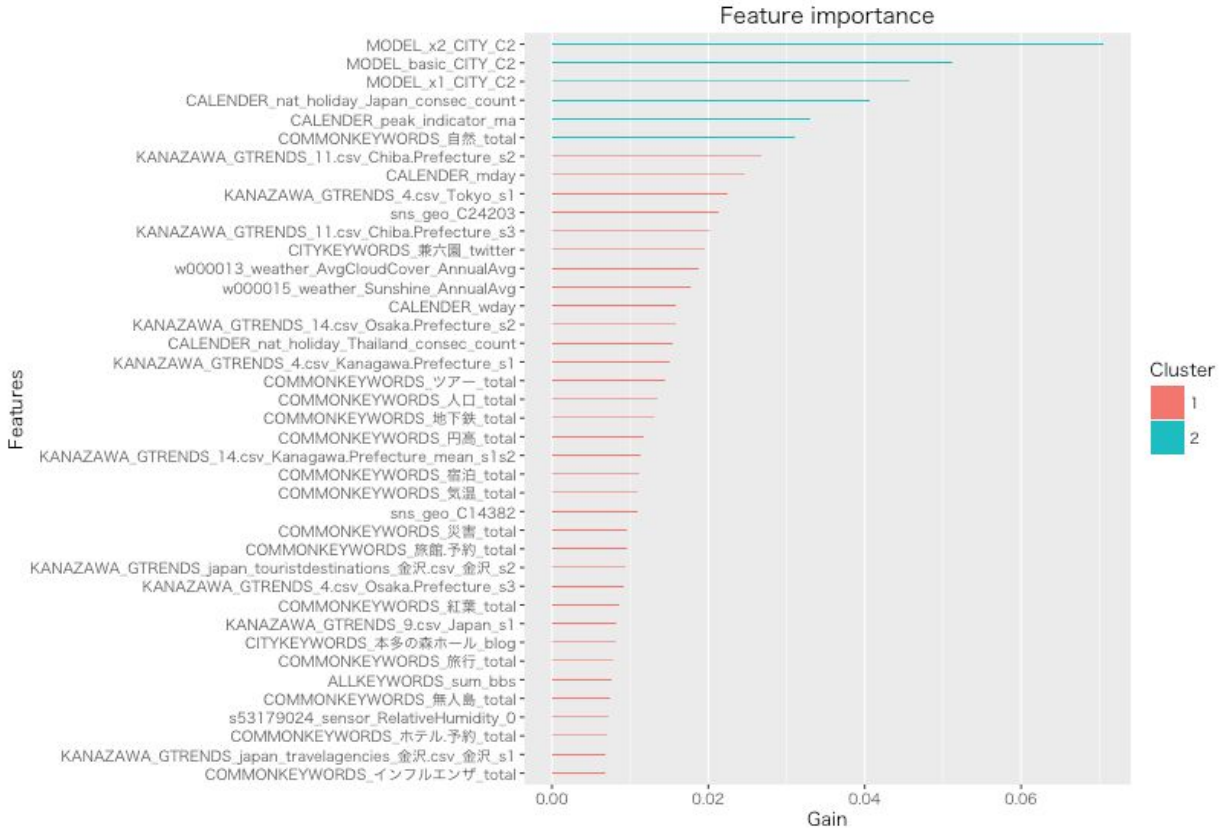
正規化した富山市(C6)の観光客数とGoogle Trends (Tokyo_Travel_富山市_3週間前)データの相似



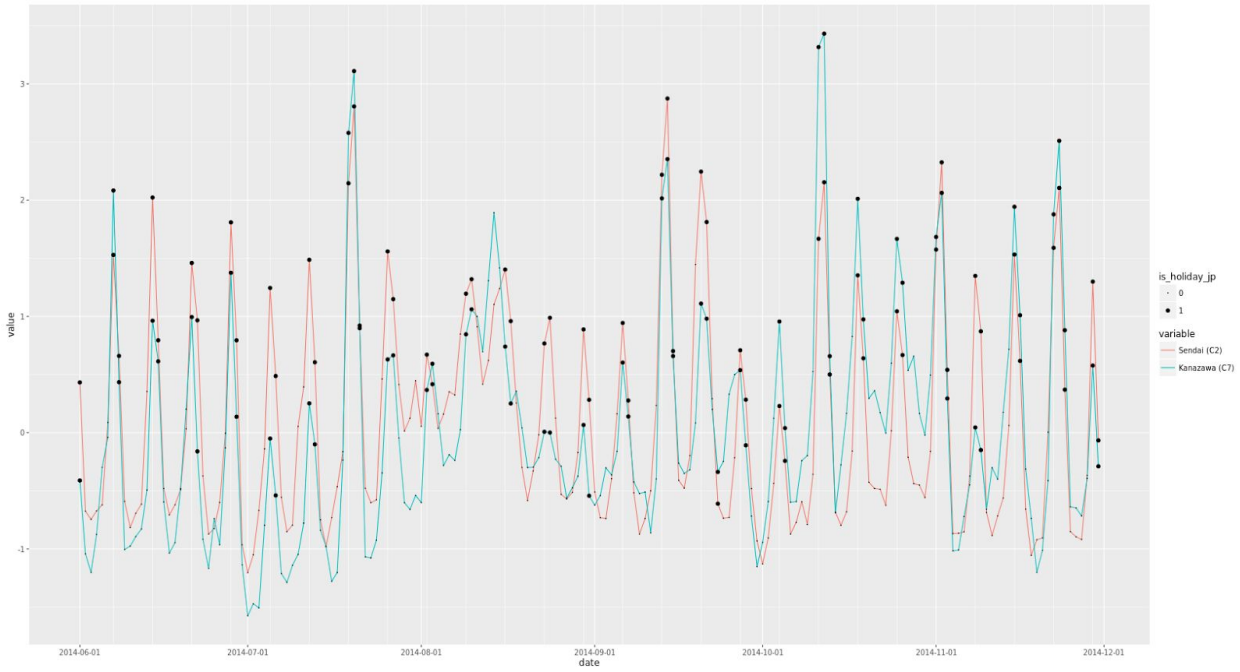
金沢市のmodel_x2における説明変数の寄与度（トップ40）



金沢市のmodel_lastにおける説明変数の寄与度（トップ40）



正規化した伊勢市(C2) と金沢市 (C7) (2014年6月- 2014年11月) の観光客数傾向の相似



H. 説明変数の解釈

富山市
富山市のmodel_basicの説明変数の寄与を見ると、
3位に'GW'、10位に'家族旅行'のSNSでの出現数が入っており、この2つのキーワードとの関連が高いことがわかる。また、6位に'連休'に関するキーワードが入っており、以上を鑑みると、**富山市への観光客は、連休を用いた家族旅行が主である**と推測される。

金沢市
1位と2位に、連休に関連する説明変数が入っていることから、まとまった休みに旅行にいく人が多いと推測される。3位には東京における金沢への陸路に関する検索数が入っており、**北陸新幹線開通に伴い、陸路の選択肢がより身近になった**ことが伺える。また5位に兼六園が入っていることから、金沢ではやはり兼六園がダントツの人気であることがわかる。

I. コード詳細

OSのバージョン、ソフトウェア、モジュール等

- OS version: Ubuntu 14.03.3 LTS on AWS r3.8xlarge EC2インスタンス
- Software: R 3.2.2 with Rstudio server
- Rパッケージ: 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パラメータ.

全てのモデルにおいて、
eta = 0.01, subsample = 0.8, seed = 23.
各都市についてのパラメータ (C2, C5, C6, C7, C9)

モデル名称	**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)

コードの実行と再現の方法

1. リポジトリ（https://github.com/aakansh9/METI-tourism-prediction）をダウンロード(コードと元のデータ)
2. RunMe.Rに記載した方法でRをインストール
3. RunMe.Rの中の、main_pathにリポジトリのパスを代入
4. RunMe.Rを実行
5. 富山市 (C6) と金沢市 (C7)の予測を含んだsubmission_ensemble.csvが生成される