AXA Data Challenge

Mahmut CAVDAR (mahmutcvdr@gmail.com), **Anh Khoa NGO HO**
(anh-khoa.ngo-ho@u-psud.fr), **Trong Bach VU** (jsbachvu@gmail.com)
M2 AIC - Paris Saclay University

Abstract

AXA data challenge is based on the economic problem of supply and demand, which means balancing the priorities of service levels and labour cost in AXA contact center. In this case, the main purpose is the accurate workload forecast for having the right number of staff in this contact center. For this challenge, an inbound call forecasting system is proposed for predicting the number of incoming calls for AXA call center in France.

1. Introduction

In AXA data challenge, the workload forecast uses the algorithm of regression providing a model with high accuracy that predicting incoming call number on a per half-hour time slot. This challenge provides a training dataset and a platform for accuracy evaluation of model proposed. The evaluation metric used for giving a higher penalty to underestimating the number of calls is LinEx loss, which shows the seriousness of falling customer satisfaction and human resource cost.

$$\text{LinEx}(y, \hat{y}) = \exp(\alpha(y - \hat{y})) - \alpha(y - \hat{y}) - 1$$

where y is the true number of call, $\hat{y}$ is the predicted number and α = -0.1. The model is built on machine learning library scikit-learn of Python and its regression algorithms are applied for finding the acceptable result.

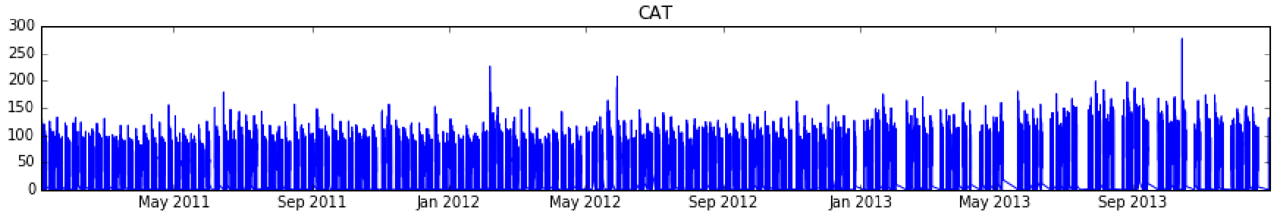
The rest of this report is organized as follows: The data is first analysed and then pre-processed by modifying its features and its values in Sect. 2. The models from regression algorithms are trained and evaluated in Sect. 3. Finally, the result received and the conclusion is in Sect. 4.

2. AXA Data Analysis and Pre-processing

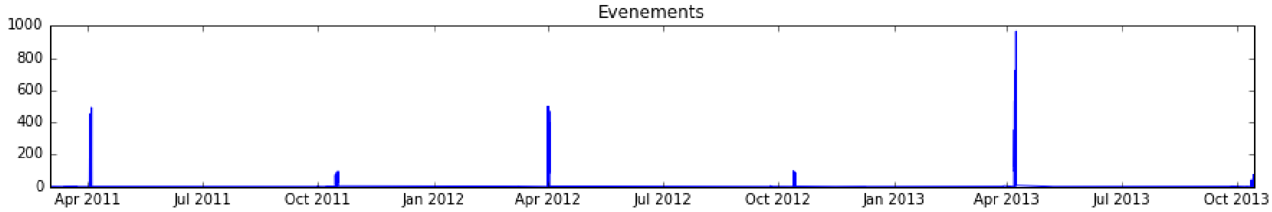
The data provided is of Comma Separated Values format with a size of nearly 3.5 gigabyte. The 10878470 incoming calls are collected from 2011 to 2013. The main features are time stamps in half-hour slots (DATE) and assignment (ASS_ASSIGNMENT), and in this data there are 87 features which show all of informations relating to these calls. The feature for prediction is number of incoming calls (CSPL_RECEIVED_CALLS). The data for evaluation is the calls of twelve weeks different and separated from December 2012 to 2013. It has two features, namely DATA and ASS_ASSIGNMENT.

As can be seen from the evaluation dataset, the number of incoming call is predicted mainly by feature time, which means that the features of training data relating to time should be kept such as DAY_OFF, WEEK_END, DAY_WE_DS, TPER_TEAM and DATE. There are features represented by a string which should be numbered. In the case of DATE, this feature is a string including year, month, day, hour, minute and second and thus it is splitted into features YEAR, MONTH, DAY, MOMENT. In the same situation, DAY_WE_DS showing weekdays is numbered from 0 to 6 and TPER_TEAM is 0 (Day) or 1(Night). As a result, the training and testing dataset are modified and the new features mentioned are added for the training process.

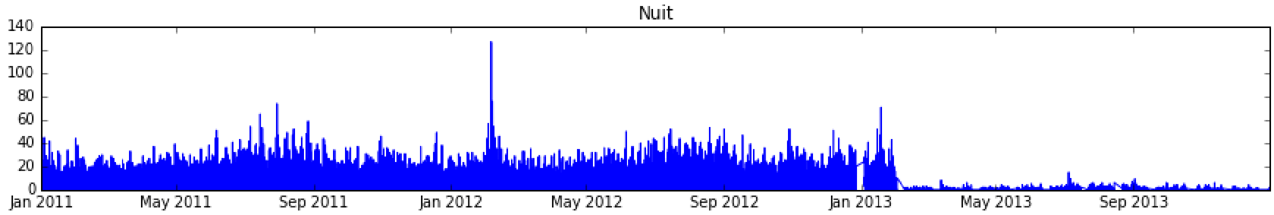
In ASS_ASSIGNMENT, there are 28 subjects that its number of calls is predicted in half-hour slots. We recognize that these calls and subjects are independent, hence we build a different model for each subject, which is based on its data changes over time. We discuss in this section remarkable assignment subjects such as CAT [2.1], Evenements [2.2], Nuit [2.3], Telephonie [2.4]. Its graphs below show the number of incoming call over the whole period from 2011 to 2013. The other graphs are found in the appendix section of this report.



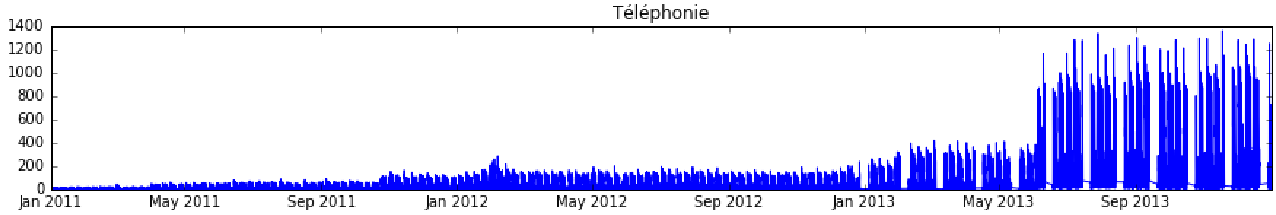
Graph 2.1: Number of call in CAT



Graph 2.2: Number of call in Evenements

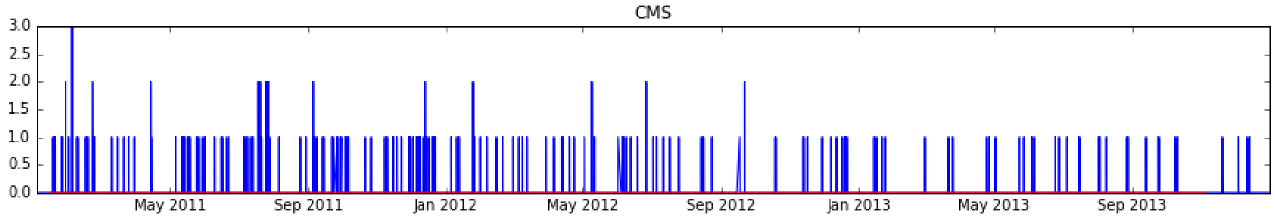


Graph 2.3: Number of call in Nuit

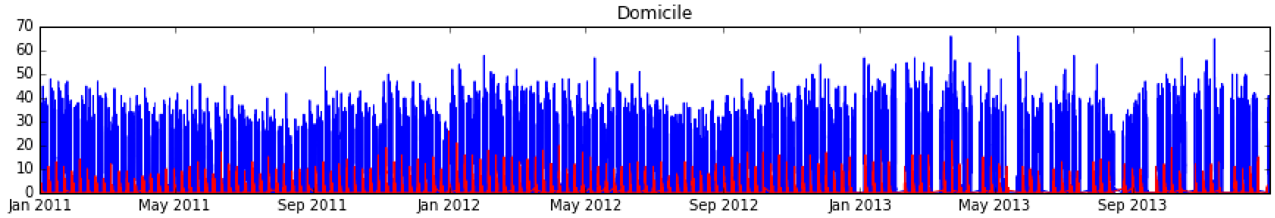


Graph 2.4: Number of call in Telephonie

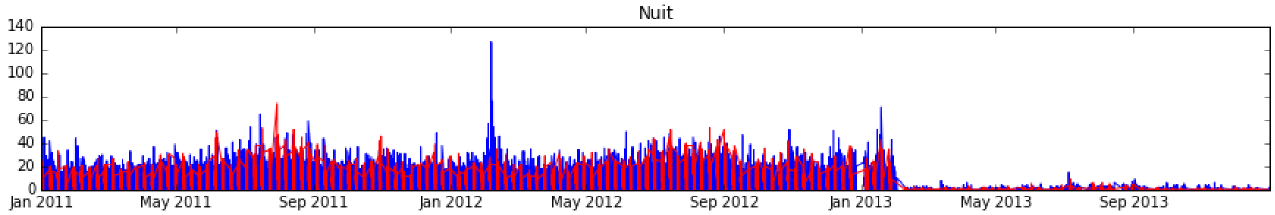
In the case of CAT, this bar graph [2.1] describes generally similar patterns from March 2011 to December 2013. In contrast, Evenements graph [2.2] has the figures significant in April and October and the rest periods have no call, which means that the calls of Evenements are only in two period of April and October. Moreover, the number of this subject ranges dramatically from 0 to more than 900 calls. For Nuit [2.3], there is a shape decrease from February 2013, followed by a slight changes in the rest of period. Telephonie has a moderate upward trend from January 2011 to May 2013 and a quick growth from June 2013. Therefore, each assignment has its own model with the different parameters and the model uses the dataset differently.



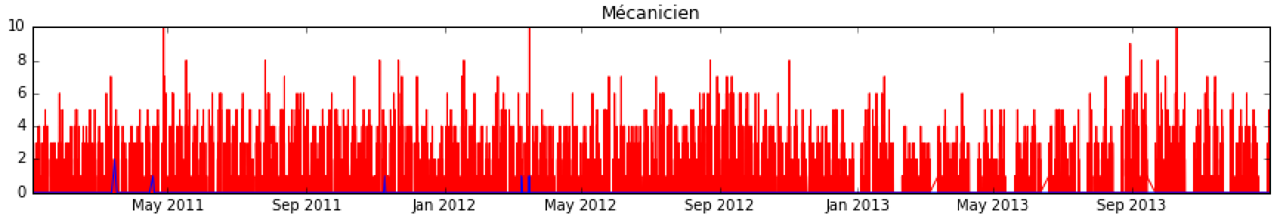
Graph 2.5: Feature WEEK_END of CMS (Blue: Weekdays, Red: Weekends)



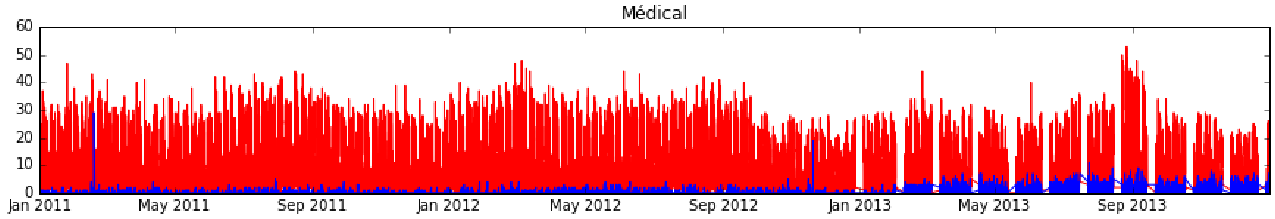
Graph 2.6: Feature WEEK_END of Domicile (Blue: Weekdays, Red: Weekends)



Graph 2.7: Feature WEEK_END of Nuit (Blue: Weekdays, Red: Weekends)



Graph 2.8: Feature TPER_TEAM of Mecanicien (Red: Day, Blue: Night)



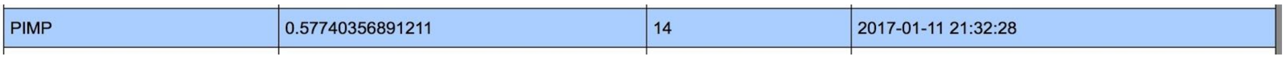
Graph 2.9: Feature TPER_TEAM of Medical (Red: Day, Blue: Night)

In feature selection matter, for some assignments, there are features presenting clearly the differences such as WEEK_END or TPER_TEAM. For instance, CMS [2.5] assignment has no incoming call in weekends but Domicile [2.6] has a number of calls in the same period. In the case of Nuit [2.7], it is more difficult in detection because of the similar patterns between weekdays and weekends. The same situation with TPER_TEAM, there are also the differences between day and night in Mecanicien [2.8] and Medical [2.9]. More details of **Feature Engineering** are in the **Report_plot.ipynb** file.

3. Learning Algorithm and Evaluation

This section discuss the learning algorithms used for predicting the number of incoming calls (CSPL_RECEIVED_CALLS). In this case, we select the regression algorithms such as Tree Regressors, Random Forest Regressors, Gradient Boosting Regressors, Autoregressive Models. After training the data with the algorithm mentioned and its default parameter, we select the most promising algorithm, namely Gradient Boosting Regressors, with the best accuracy for the next process. We then modify its parameters such as the number of boosting stages, the maximum depth of the individual regression estimators, the minimum number of samples required to split an internal node, the learning rate and the loss function. These parameters are different for each assignment, which means that there are 27*52 models used for the prediction (for each different group and each different period). Moreover, for each period in the testing dataset, we use different datasets. In detail, there are 52 periods discontinued which are the testing dataset. As a result, the loss value is significantly decreased from 2000 to 500 after many submissions.

As can be seen from the previous part, based on the changes of incoming call number of each assignment from 2011 to 2013, we re-evaluate the result of the models built. Therefore, our final loss value is decreased to 0.577. We utilize 1.535 coefficient after modeling phase. It’s an experimental constant.



PIMP	0.57740356891211	14	2017-01-11 21:32:28

Image 3.1: Leaderboard - LinEx Loss

4. Conclusion

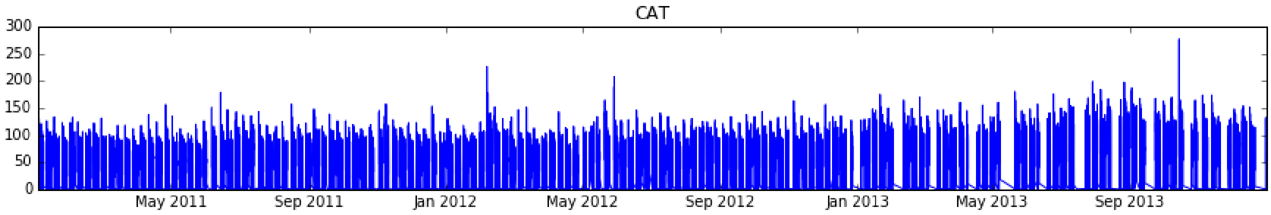
In this paper, we proposed a set of models for inbound call forecasting system with the algorithm Gradient Boosting Regressor. In fact, each model built is based on the pattern of each assignment data from 2011 to 2013, which reduces finally the LinEx loss to 0.577. Therefore, this system could be applied in predicting the number of incoming calls for AXA call center.

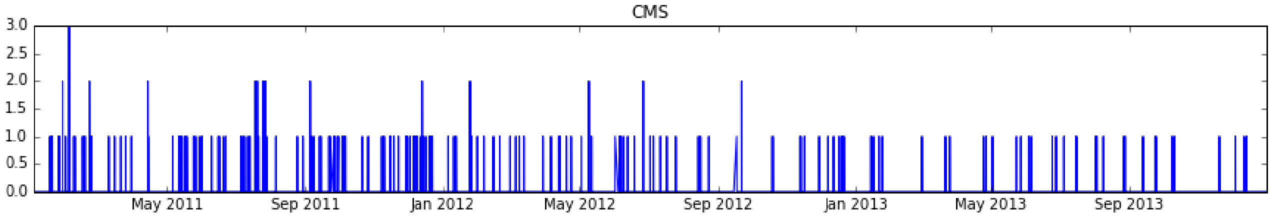
Reference

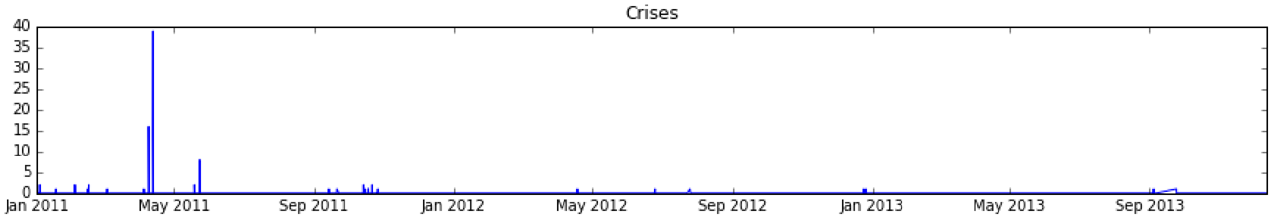
[1] Python - https://www.python.org

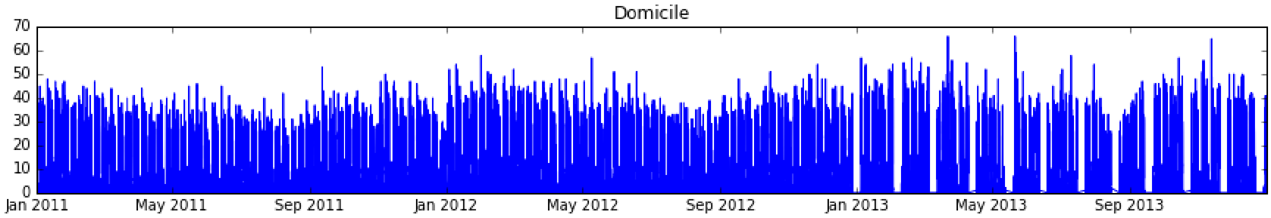
[2] Scikit-learn - http://scikit-learn.org/

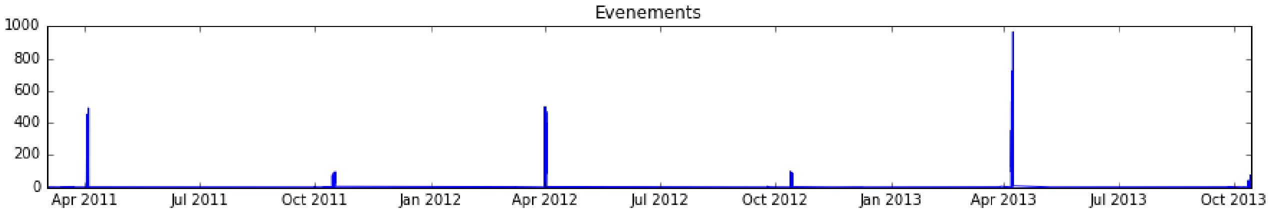
Appendix

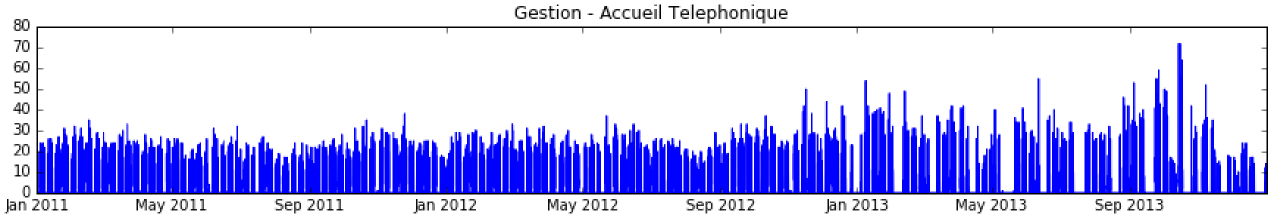

CAT
300
250
200
150
100
50
0
May 2011
Sep 2011
Jan 2012
May 2012
Sep 2012
Jan 2013
May 2013
Sep 2013

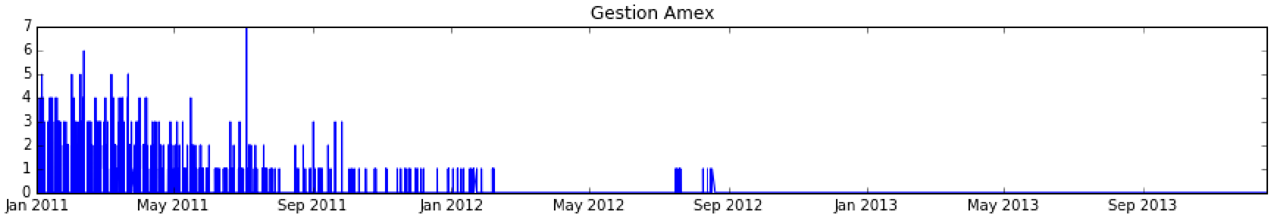

CMS
3.0
2.5
2.0
1.5
1.0
0.5
0.0
May 2011
Sep 2011
Jan 2012
May 2012
Sep 2012
Jan 2013
May 2013
Sep 2013

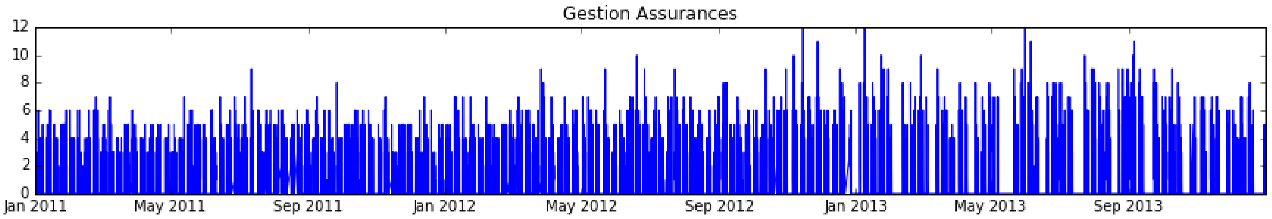

Crises
40
35
30
25
20
15
10
5
0
Jan 2011
May 2011
Sep 2011
Jan 2012
May 2012
Sep 2012
Jan 2013
May 2013
Sep 2013

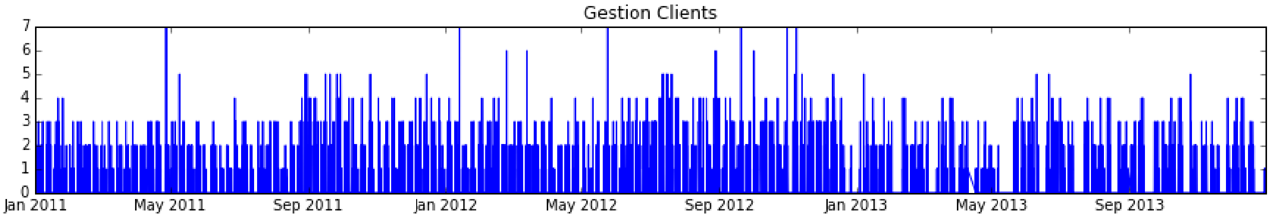

Domicile
70
60
50
40
30
20
10
0
Jan 2011
May 2011
Sep 2011
Jan 2012
May 2012
Sep 2012
Jan 2013
May 2013
Sep 2013

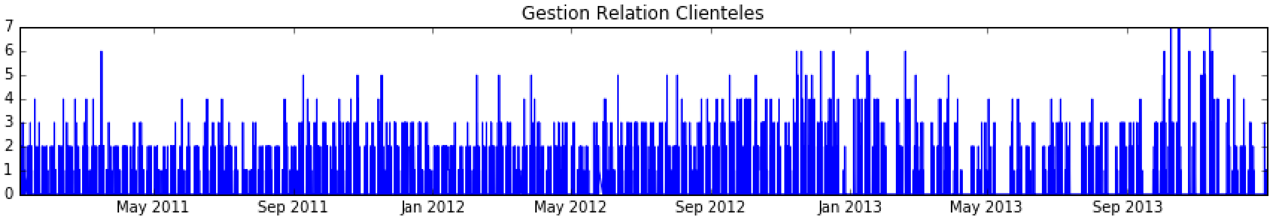

Evenements
1000
800
600
400
200
0
Apr 2011
Jul 2011
Oct 2011
Jan 2012
Apr 2012
Jul 2012
Oct 2012
Jan 2013
Apr 2013
Jul 2013
Oct 2013

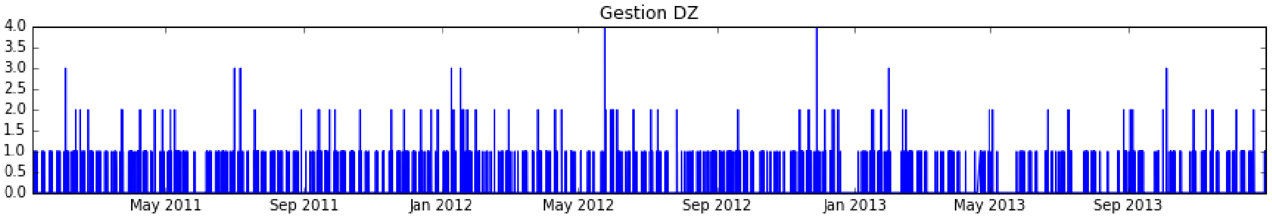

Gestion - Accueil Telephonique
80
70
60
50
40
30
20
10
0
Jan 2011
May 2011
Sep 2011
Jan 2012
May 2012
Sep 2012
Jan 2013
May 2013
Sep 2013

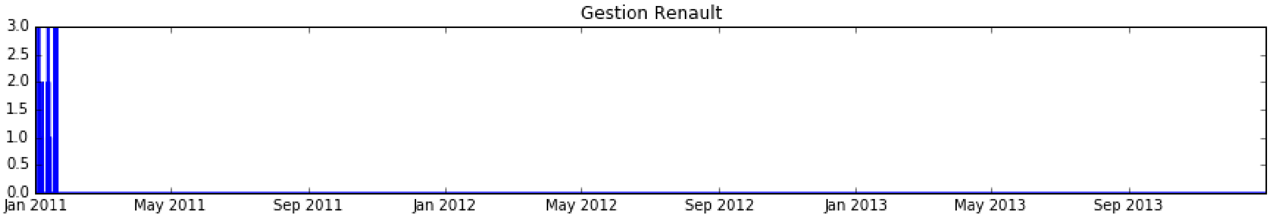

Gestion Amex
7
6
5
4
3
2
1
0
Jan 2011
May 2011
Sep 2011
Jan 2012
May 2012
Sep 2012
Jan 2013
May 2013
Sep 2013

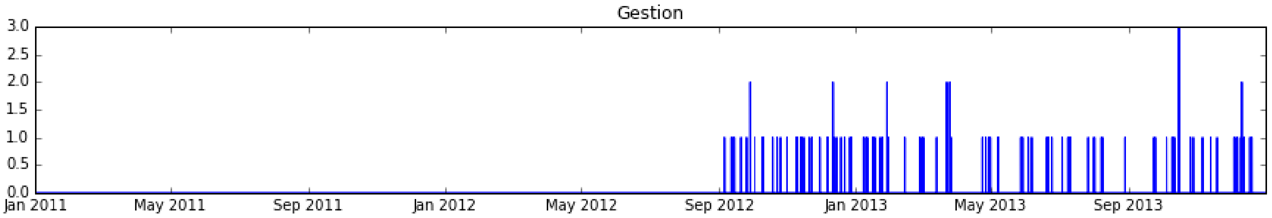

Gestion Assurances
12
10
8
6
4
2
0
Jan 2011
May 2011
Sep 2011
Jan 2012
May 2012
Sep 2012
Jan 2013
May 2013
Sep 2013

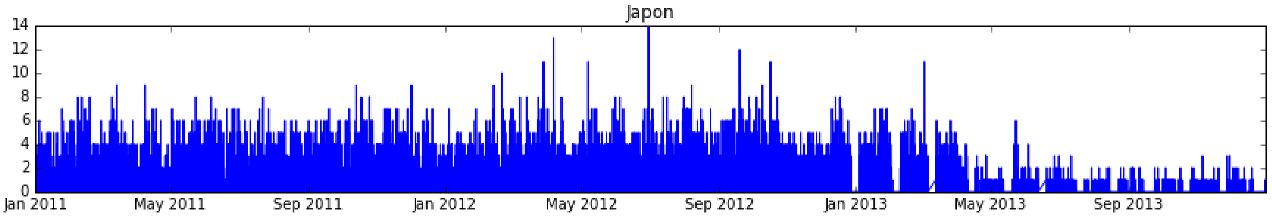

Gestion Clients
7
6
5
4
3
2
1
0
Jan 2011
May 2011
Sep 2011
Jan 2012
May 2012
Sep 2012
Jan 2013
May 2013
Sep 2013

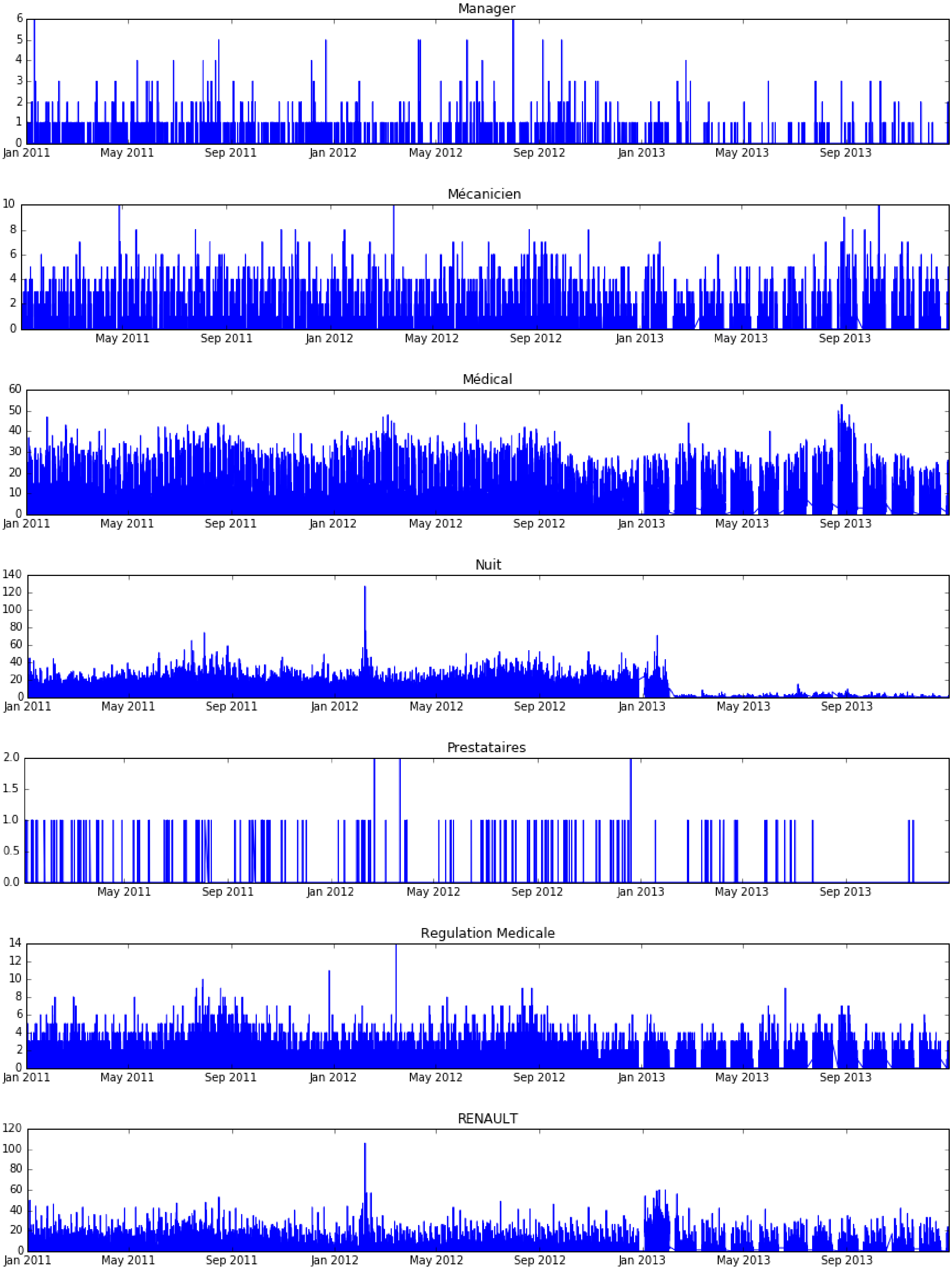

Gestion Relation Clienteles
7
6
5
4
3
2
1
0
May 2011
Sep 2011
Jan 2012
May 2012
Sep 2012
Jan 2013
May 2013
Sep 2013

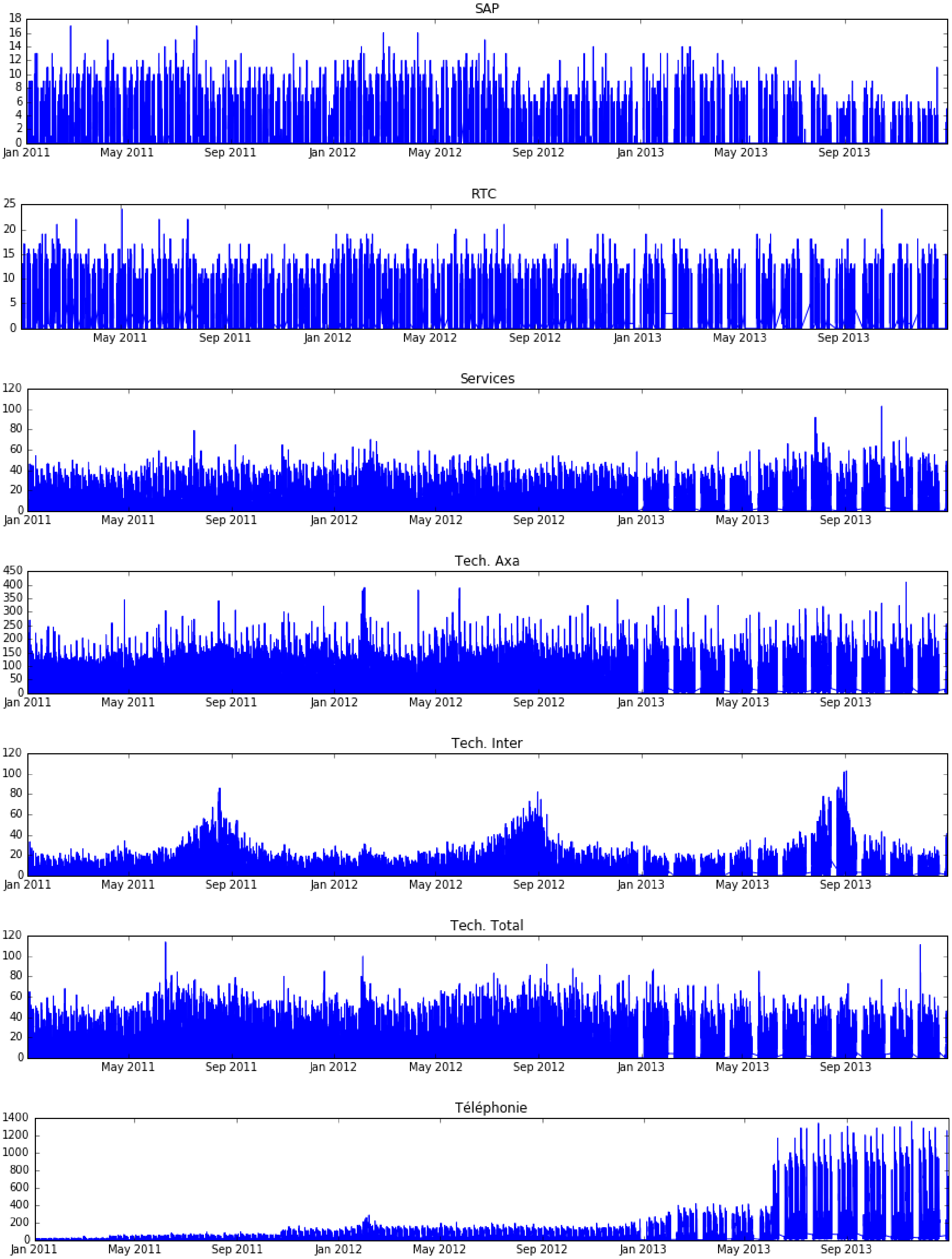

Gestion DZ
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
May 2011
Sep 2011
Jan 2012
May 2012
Sep 2012
Jan 2013
May 2013
Sep 2013


Gestion Renault
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Jan 2011
May 2011
Sep 2011
Jan 2012
May 2012
Sep 2012
Jan 2013
May 2013
Sep 2013


Gestion
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Jan 2011
May 2011
Sep 2011
Jan 2012
May 2012
Sep 2012
Jan 2013
May 2013
Sep 2013


Japon
14
12
10
8
6
4
2
0
Jan 2011
May 2011
Sep 2011
Jan 2012
May 2012
Sep 2012
Jan 2013
May 2013
Sep 2013


Manager
0
1
2
3
4
5
6
Jan 2011
May 2011
Sep 2011
Jan 2012
May 2012
Sep 2012
Jan 2013
May 2013
Sep 2013
Mécanicien
0
2
4
6
8
10
May 2011
Sep 2011
Jan 2012
May 2012
Sep 2012
Jan 2013
May 2013
Sep 2013
Médical
0
10
20
30
40
50
60
Jan 2011
May 2011
Sep 2011
Jan 2012
May 2012
Sep 2012
Jan 2013
May 2013
Sep 2013
Nuit
0
20
40
60
80
100
120
140
Jan 2011
May 2011
Sep 2011
Jan 2012
May 2012
Sep 2012
Jan 2013
May 2013
Sep 2013
Prestataires
0.0
0.5
1.0
1.5
2.0
May 2011
Sep 2011
Jan 2012
May 2012
Sep 2012
Jan 2013
May 2013
Sep 2013
Regulation Medicale
0
2
4
6
8
10
12
14
Jan 2011
May 2011
Sep 2011
Jan 2012
May 2012
Sep 2012
Jan 2013
May 2013
Sep 2013
RENAULT
0
20
40
60
80
100
120
Jan 2011
May 2011
Sep 2011
Jan 2012
May 2012
Sep 2012
Jan 2013
May 2013
Sep 2013


SAP
0
2
4
6
8
10
12
14
16
18
Jan 2011
May 2011
Sep 2011
Jan 2012
May 2012
Sep 2012
Jan 2013
May 2013
Sep 2013
RTC
0
5
10
15
20
25
May 2011
Sep 2011
Jan 2012
May 2012
Sep 2012
Jan 2013
May 2013
Sep 2013
Services
0
20
40
60
80
100
120
Jan 2011
May 2011
Sep 2011
Jan 2012
May 2012
Sep 2012
Jan 2013
May 2013
Sep 2013
Tech. Axa
0
50
100
150
200
250
300
350
400
450
Jan 2011
May 2011
Sep 2011
Jan 2012
May 2012
Sep 2012
Jan 2013
May 2013
Sep 2013
Tech. Inter
0
20
40
60
80
100
120
Jan 2011
May 2011
Sep 2011
Jan 2012
May 2012
Sep 2012
Jan 2013
May 2013
Sep 2013
Tech. Total
0
20
40
60
80
100
120
May 2011
Sep 2011
Jan 2012
May 2012
Sep 2012
Jan 2013
May 2013
Sep 2013
Téléphonie
0
200
400
600
800
1000
1200
1400
Jan 2011
May 2011
Sep 2011
Jan 2012
May 2012
Sep 2012
Jan 2013
May 2013
Sep 2013