Social Analysis of Fuorisalone 2016

Machine Learning Applied to Affluence Prediction

Authors:

Riccardo DE TOGNI {riccardo.detogni@mail.polimi.it}

Andrea DONATI {andrea4.donati@mail.polimi.it}

Nicola GHIO {nicola.ghio@mail.polimi.it}

Prof: Marco Brambilla

Milan, January 12, 2018

Contents

1 Introduction

Social analysis is the practice of systematically examining a social problem, issue or trend, often with the aim of capture changes in the situation being analyzed. In this document we describe from scratch the entire process of analyzing a big event such as Fuorisalone, starting from data collected in different ways that have to be aggregated and integrated, ending with a machine learning prediction model.

2 Problem Definition

The aim of this project is to perform an analysis on data from Fuorisalone 2016 in order to find any interesting correlation or pattern and to try to predict the affluence to single events that are part of it. During Fuorisalone week many firms organize different kind of events, from design exhibitions to night parties, all over Milan. This heterogeneity is measured in terms of different categories of events, their locations, duration and time of the day in which they were active. It is possible to perform many kind of analysis in order to find interesting patterns, in particular we mainly focused on finding the macrozones of the events and the most popular ones, the most relevant categories of events and the most popular ones. We also decided to create a regression model to predict the affluence to the events.

3 Data Exploration

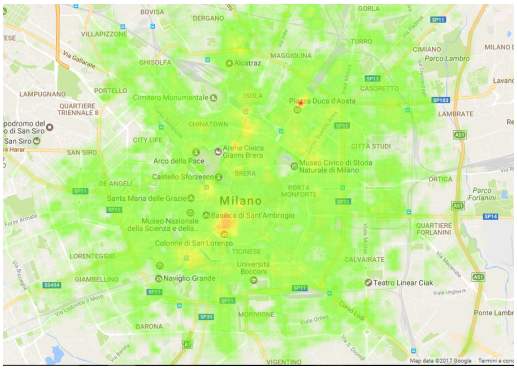
In this section are discussed the datasets we used for our analysis. We will explain how they are composed, and the analysis we performed on it. Since many of fuorisalone events lasts during the whole week, we decided to summarize the days, considering 4 aggregated periods of 6 hours that we called timeslots.

3.1 Phone Activities Dataset

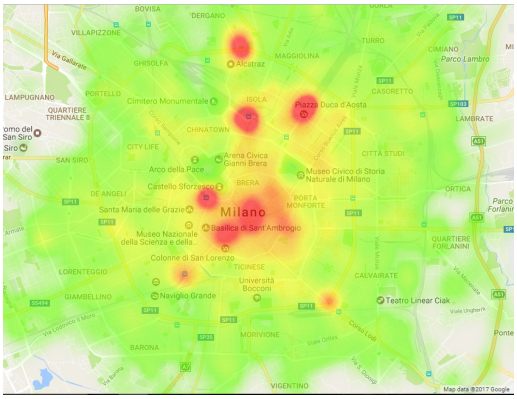
By concession of Telecom Italia, it is composed of 7.242.109 tuples with the following attributes. An activity corresponds to a single action performed by a user, for example it could be a phone call or an sms.

- Cell ID
- Age Range: [<18] [18-30] [31-40] [41-60] [51-60] [>60]
- Contract type: Private or Corporate
- Gender: Male or Female
- Day
- Timeslot: divided in [00.00 - 6.00] [6.00-12.00] [12.00 - 18.00] [18.00 - 24.00]
- Latitude and longitude
- Number of activities

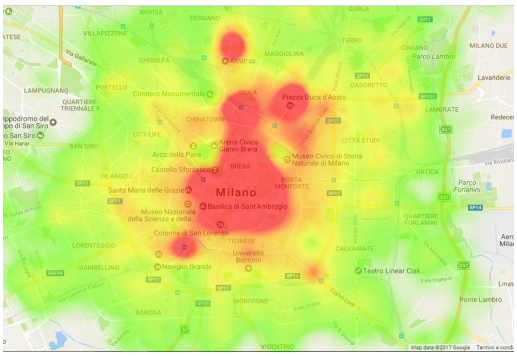
In order to understand a general trend in cell activity we plot a heatmap representing the density of usage among all the cells. Each map stands for a different Timeslot. As we expect the activity is rarefied between midnight and 6 a.m. and it incrementally increase until 18 p.m..



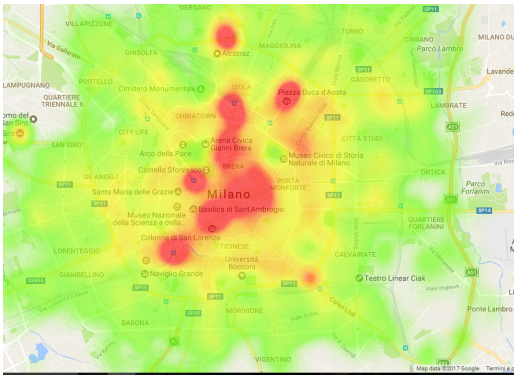
(a) From 0 a.m. to 6 a.m.



(b) From 6 a.m. to 12 a.m.



(a) From 12 a.m. to 6 p.m.



(b) From 6 p.m. to 12 p.m.

Figure 2: Phone activities during in different time slots during the day

3.2 Fuorisalone Event Data

Fuorisalone App contains all the info about events that took place in Milan during the Fuorisalone week in 2016. We extracted it from the app database together with their locations and we joined them in order to create the event dataset. The most important features of the datasets are:

- Categories
- Location
- Date and Time

All the other features like the event id, name, address, etc. have been discarded as not relevant for our problem.
After that we analyzed the geographical distribution of the events in the city and, by plotting them, we can clearly see that some areas are more important than others, since the concentration of events is higher.

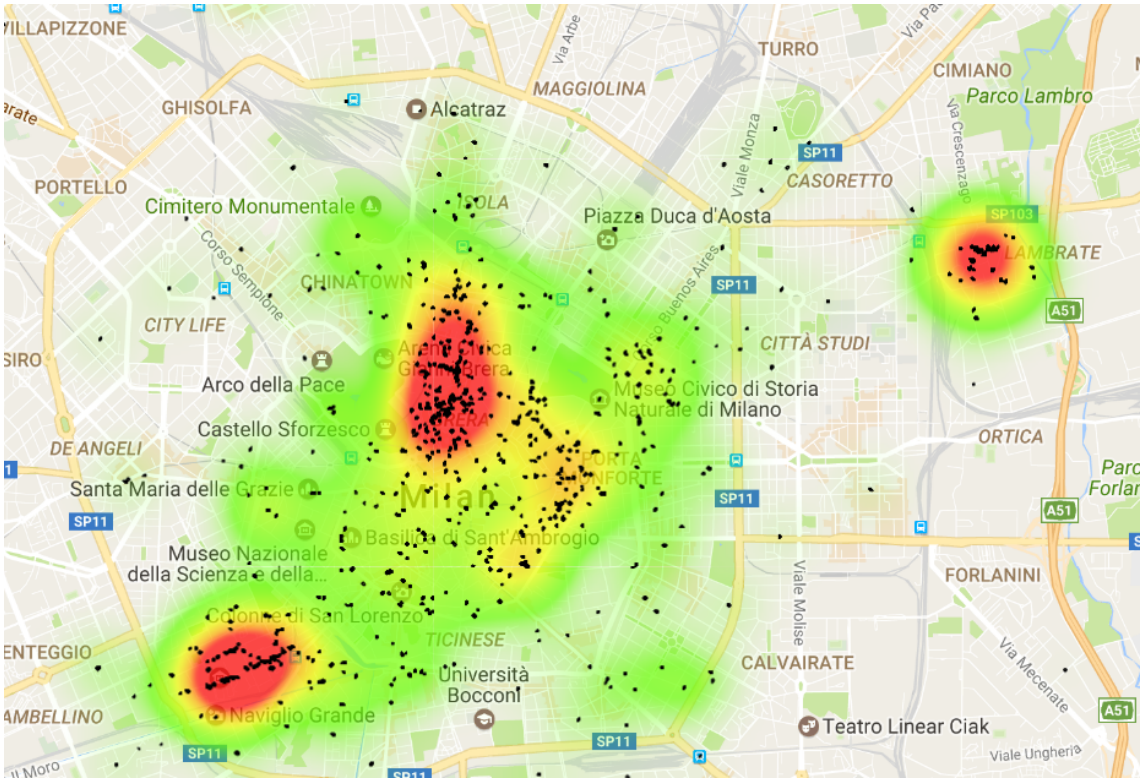
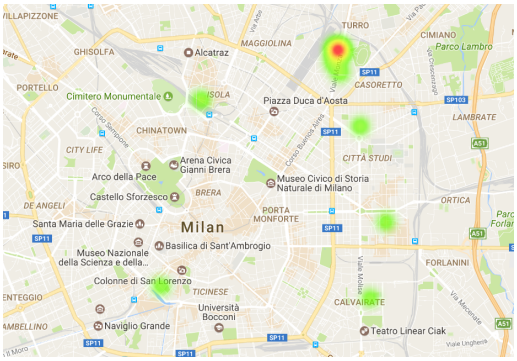


Figure 3: Geographical distribution of events in Milan

In particular we can see that the most crowded areas are: Brera, Tortona, Lambrate and, to some extent, the city center around Duomo. From this visualization we could have divided the events into zones by hand, but we decided to go for a more automatic way using density based clustering, whose results are presented in section 4.2.

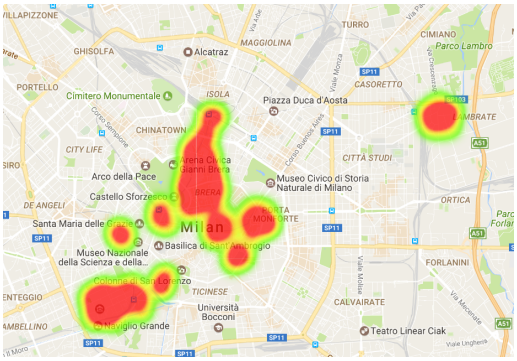
3.3 Fuorisalone App Positions

Fuorisalone App registers the position of users randomly picked when they are actually using the app on their smartphone. In particular, we can count on about 300k positions spread all over April 2016 (250k during Fuorisalone week). Plotting a heatmap of all the positions gives an idea of which are the most frequented areas: it is not surprising that these areas are Tortona, Brera, Lambrate and generally the city center. But we go deeper in the analysis of position data, we tried to exploit a possible trend of people movements by clustering positions during different timeslots along the whole week. As we can see in the first timeslot, from midnight to 6 a.m., there is barely no activity: small clusters in movida areas. The second timeslot starts to show the main areas of Fuorisalone. The third timeslot, 12 a.m. to 6 p.m. while keeping the main areas showed before, it highlights two big new areas: Piazza Duomo and San Babila. The fourth timeslot represent the position between 6 p.m. and midnight. As we can see probably people move to areas with Restaurants and clubs like Tortona, Navigli and Brera.



(a) From 0 a.m. to 6 a.m.

(b) From 6 a.m. to 12 a.m.



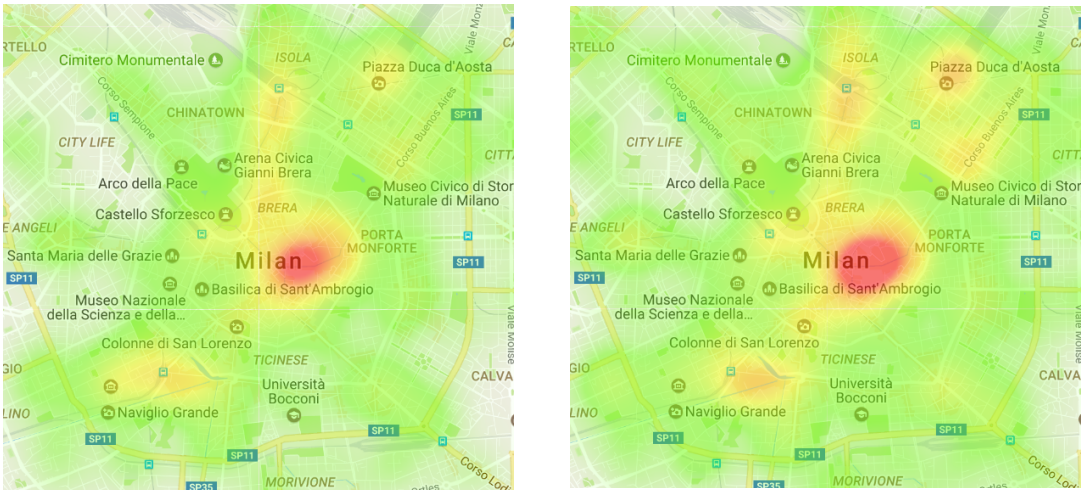
(a) From 12 a.m. to 18 p.m.

(b) From 18 p.m. to 24 p.m.

Figure 5: App positions in different time slots during the day

3.4 Foursquare Dataset

The FourSquare Dataset is composed by check-ins made during April 2016 in Milano metropolitan area. Since FourSquare is not widely adopted in Italy, the dataset counts only a little more than 91k records. If we slice it to keep the check-ins made during Fuorisalone week the counter goes down to 21k.



(a) Check-in during Fuorisalone

(b) Check-in during whole month

Figure 6: FourSquare Dataset April 2016

As we can see in the pictures, the trend is almost the same during the festival and during the rest of the month. Furthermore, we see that the largest part of check-ins is made near Piazza Duomo, Stazione Centrale and Stazione Garibaldi. This lead us to the conclusion that FourSquare Dataset would not bring us any interesting information about the physical presence of people at FuoriSalone. Considering also the small number of records we decided to discard FourSquare Dataset from our analysis.

4 Data Preparation

4.1 Creation of the Datasets

The definitive dataset used for the prediction is composed by events, described by input features like "Categories", "Timeslots", "Geographical Cluster" and output features: Position Count, Activity Count. In order to obtain it we performed a complex integration among different datasets. The Geographical cluster is assigned to each event with a DBSCAN clustering model, minimizing the number of outliers events (event which do not belong to any cluster). To estimate the affluence of people to an event we assume that positions in a radius of 150 meters from an event are indeed attending that event. Similarly we assign to each event a phone cell taken from Phone Dataset, described above. From the cell data we computed the correspondent Activity Count, differentiated by Timeslot. The resulting dataset, after the encoding of categorical attributes, counts 55 features. In order to understand if the two Output variable we selected were someway connected we compute a spurious correlation for the main event clusters. The result are plotted in figure 9, where we can see a strong correlation between Activity and Position count for all the selected clusters.

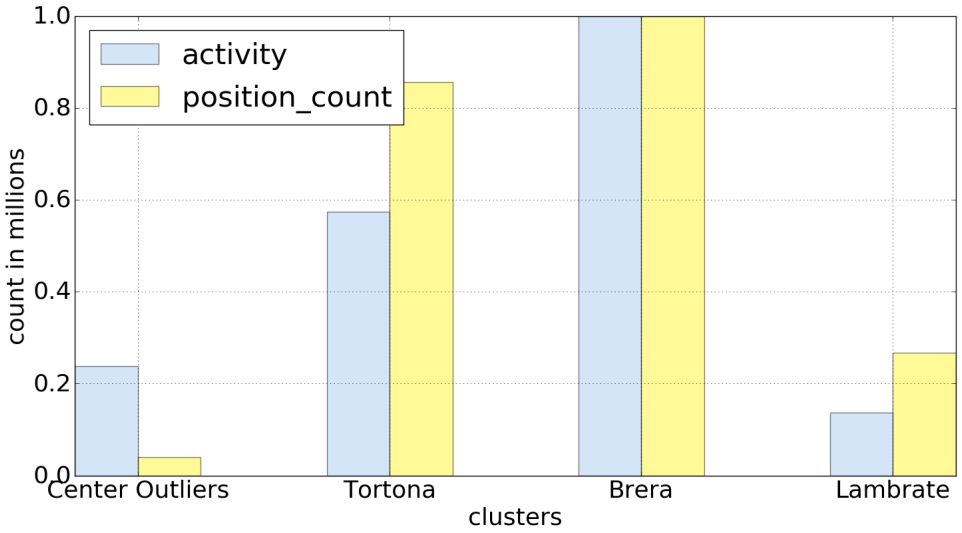


Figure 7: Phone activities and app positions for different clusters

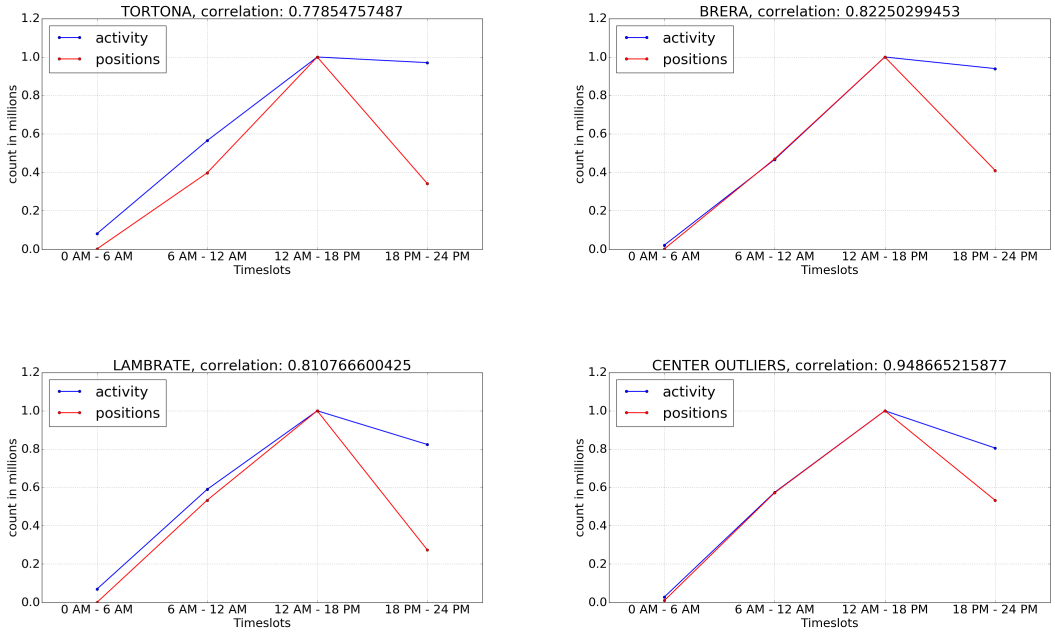


Figure 9: Trend of phone activities and positions during the day for different clusters

4.2 Clustering of events

As mentioned above, we clustered events geographically. In this way , we will be able to use geographical information of new events as further useful information for prediction. The basic idea behind our clustering is maximize coverage and at the same time minimize the number of cluster. The procedure is made of two steps: first we clustered with DBSCAN finding a dozen of clusters,

then we clustered again just the outliers of the first step. The result is a 31 clusters division with about 90% of coverage.

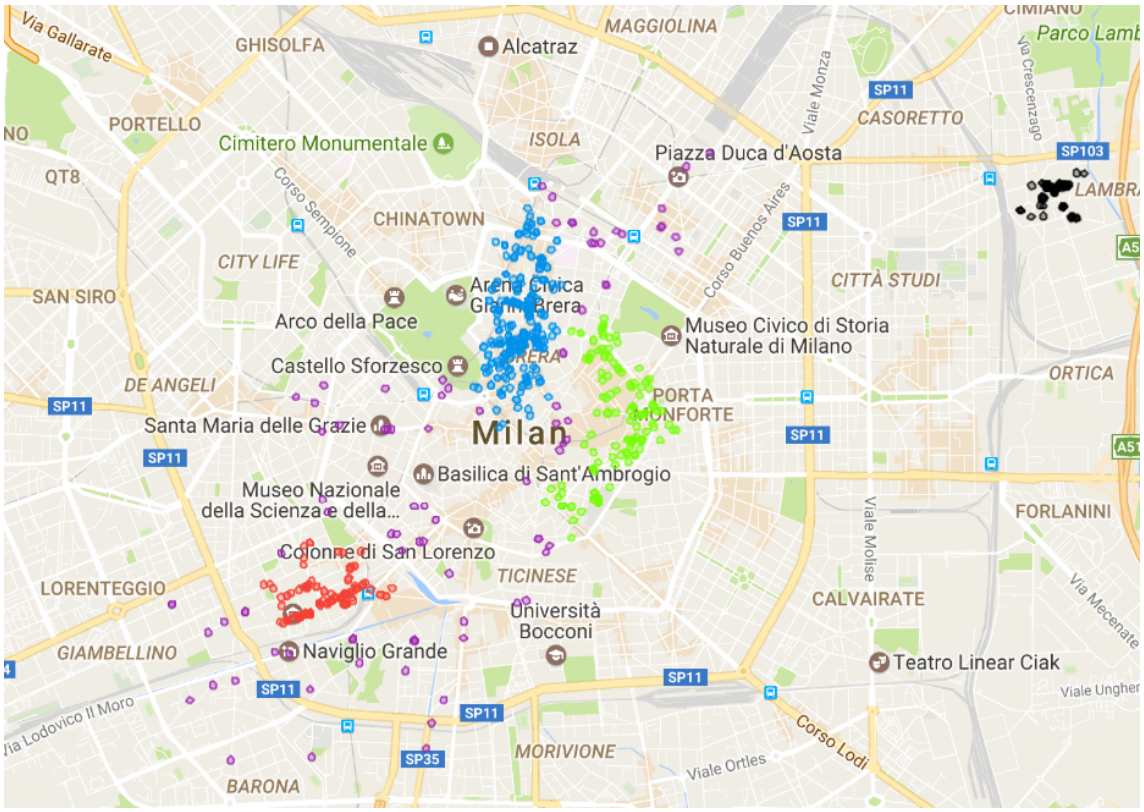


Figure 10: Main Clusters

Figure 10 is a plot of 4 main clusters, Tortona in red, Lambrate in black, Brera in blue and city center/Statale Milano in green. The purple points represent the biggest cluster obtained from outliers. It could represent events that are quite in city center but out of main areas. It is an interesting pattern as we will see in feature selection.

4.3 Dimensionality Reduction

Since the dataset includes 55 features we needed to reduce that number by limiting the loss of information.
The first attempt of dimensionality reduction was Principal Component Analysis (PCA), that is a form of unsupervised dimensionality reduction because it doesn't look at the target feature. By using PCA it's possible to create new dimensions that are linear combinations of original ones at the cost of a loss of represented variance of the original data.
In our specific case, we found out that it was possible to drastically reduce the number of dimensions from 55 to 24, saving the 90% of the original variance of the data.
This was only the first attempt of dimensionality reduction and, as we'll explain in section 5.1, it wasn't the one we chose in the end for the training of the model. In fact, by supervising feature selection, we obtained around the same number of features than using PCA, but the performances of some models are significantly better. Another advantage of supervised feature selection is that is more readable, in fact it can give us important information about which features are more relevant

in the prediction of the output.

5 Prediction

5.1 Feature Selection

Feature Selection represent a fundamental step in Machine Learning processes. As we have seen above, the dataset has 55 features. The aim of feature selection is to remove some of them limiting the loss of performance and keeping the variance of input data. We used a recursive algorithm called Recursive Feature Elimination with Cross Validation (RFECV). It fits the model for every possible subset of features and take the one with higher mean score among the folds, computed with R2 metric. We applied this algorithm to every model, both for Position Count and Phone Activity. As for some model we do not improve the performances, for some others we drastically enhance the R^2 score. The test made and the result are collected in the table below.

Table 1: R^2 scores for phone activities

R^2	**Activities**		
	Full Dataset	**Feature Selected**	**T-test p-value**
Ridge	0.40837 ± 0.21322	0.41296 ± 0.22714	0.93049
Bayesian Regression	0.40462 ± 0.21163	0.41290 ± 0.22781	0.87490
Random Forest	0.07505 ± 0.73556	0.29368 ± 0.49605	0.15653
Lasso	0.40819 ± 0.21337	0.41052 ± 0.23071	0.96508

Table 2: R^2 scores for app positions

R^2	**Positions**		
	Full Dataset	**Feature Selected**	**T-test p-value**
Ridge	0.54091 ± 0.06231	0.54381 ± 0.06047	0.84315
Bayesian Regression	0.54096 ± 0.06120	0.54382 ± 0.06025	0.84419
Random Forest	0.58609 ± 0.15639	0.65426 ± 0.10336	**0.04260**
Lasso	0.54068 ± 0.06073	0.54190 ± 0.06012	0.93266

The resulting extracted features, regarding Position Count prediction, are surprisingly the same for all the considered models: Living, Moda, Food & Beverage, Tecnologia as event categories, some of the biggest clusters, and the 3 possible Timeslots. This suggests us that these are probably the most related features to Position Count. We observe a huge enhance, brought by feature selection, for Random Forest applied to Activity prediction. Before RFE the model was actually unable to predict (R^2 score slightly greater than zero), after the selection R^2 score jumps to 0.30. Still, the high deviation suggests that it is probably a bad predictor anyway. Random Forest works fine for position instead. We can see that it carries the best R^2 average score. It eventually do better

with feature selection. Indeed, the small p-value suggest that the difference between the two, Full dataset score and feature selection score, is statistically significant, i.e. the improvement is not due to chance.

5.2 Model Selection

As a result of the previous analysis we can clearly see that there's not a significant difference between the different models for the prediction of the phone activities, so the choice is not so determinant.
On the other hand, for what concern the prediction of the positions, we can note that the Random Forest performs significantly better than other models, and it also significantly improves the prediction quality by using the reduced dataset obtained with feature selection.
All the performance analysis were made using cross validation, so leveraging only the training set, in order to not bias the prediction error estimate given by the test set.
At the end, for the prediction of positions using the Random Forest with the new features, we obtain the following estimate of the R^2 score on the test set

$$R^2 = 0.61375$$

6 Conclusions and Results

At the end of our analysis we found out that the positions of the users given by the app were way more useful than phone activities for prediction of the affluence to events. Probably this is due to the fact that the positions are only relative to the Fuorisalone activities, while the phone activities capture a more wide set of phenomenons happening in the city.
It is important to highlight that the position count is just an indicator of the actual affluence of people. It allows to compare different event to understand which one will have higher affluence, but cannot say anything about the real number of people attending a specific event.

6.1 Further Developments

Data collected for Fuorisalone are huge. We just scratched the tip of the iceberg of all possible analysis. A very good point would be model a Recommender System for Fuorisalone App, that recommends Events to users based on previous interactions extrapolated from users' agenda.

Another interesting analysis would be to check how many of the participations of the users in their app agenda will actually become true participations at events.