|  |
| --- |
|  |

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  | |  |
| April 2015 | |  |
| Walmart Recruiting - Sales in stormy weather | |  |
| Notes | |  |
|  | |  |
| Isaac | |  |
| Authored by: Gino Tesei | |  |
|  | |  |

Exploratory analysis

# weather.csv

Exploratory analysis has been performed and related results are shown in the Excel document [weather\_elab.xlsx](https://github.com/gtesei/fast-furious/blob/master/competitions/walmart-recruiting-sales-in-stormy-weather/weather_elab.xlsx)

There’s only 2.5% of <units> field in training data with a value > 0, i.e. 97.5% of <date,store,item> in the training are 0s.

Weather imputation models

Discarded input variables

* station\_nbr
* date

All predictors are assumed as numeric (no factors)

**Models & Performance**

* Performed **basic** imputation with **BlackGuido** on Mode/Average/LineraReg and observed mean imputing performance (RMSE) **17.9**
* Performed **full** imputation with **BlackGuido** on Mode/Average/LineraReg/KNN\_Reg/PLS\_Reg/Ridge\_Reg/SVM\_Reg/Cubist\_Reg and observed mean imputing performance (RMSE) …

Predictive model #1 – basic

* For each date <d> and for each item <i> sold/predicted to be sold in the store <s>, the related train/test set has been built with (imputed) weather data of the station <st> associated to the store <s> in the key.csv file, and the related output training variable are the units sold in the train.csv file for <d,s,i>
* Feature selection: …
* Model: …