MRR Project Group 37: Sao Paulo traffic - Variable selection

Frederick Deny, Jean-Baptiste Skutnik, Lounès Moumou

23/11/2019

Baseline

For the baseline, we chose to use a stepwise variable selection, as it is a basic and general way of selecting variables. We used the stepAIC() method and got these results:

```
##       RMSE   Rsquare
## 1 3.176067 0.5944588
```

Theoritical study

It is necessary to build upon our baseline to find a better method to have more accurate predictions.

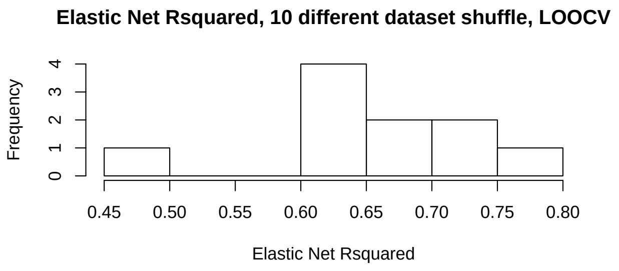
We first discarded the group-lasso method, as it requires to build groups of variables, to compute which are the most relevant. In our case, all our variables are similar (all are number of occurrences of events with the same radius of effect and range of values) so building groups would be artificial and would lack theoritical meaning.

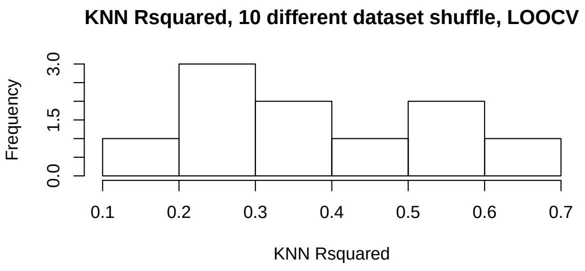
Otherwise, we believed the elastic-net to be one of our most versatile tools and chose to use it, but we mainly believed in the KNN method. Indeed, with our peculiar dataset with small ranges of values, which all supposedly induce a higher target value, the use of distance felt especially adequate.

Concerning the cross-validation method, we used the K-fold but also the LOOCV, due to the fact that some of our variables have rare occurences: in a K-fold method, these rare occurences might only be met in the testing set.

Practical study

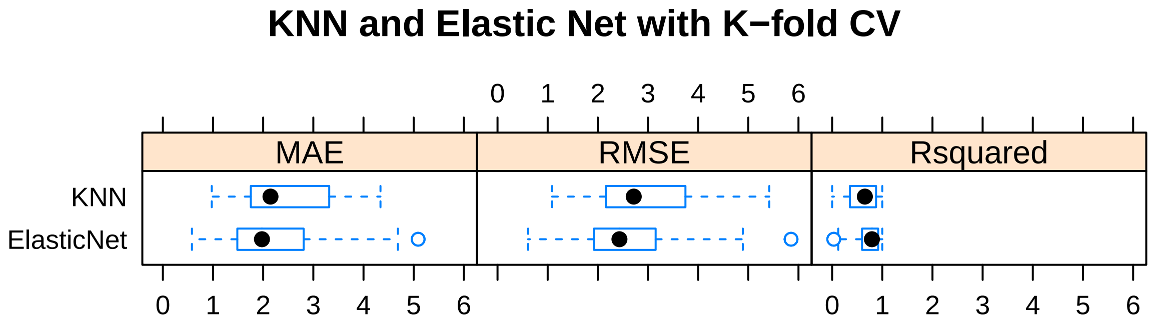
LOOCV Method

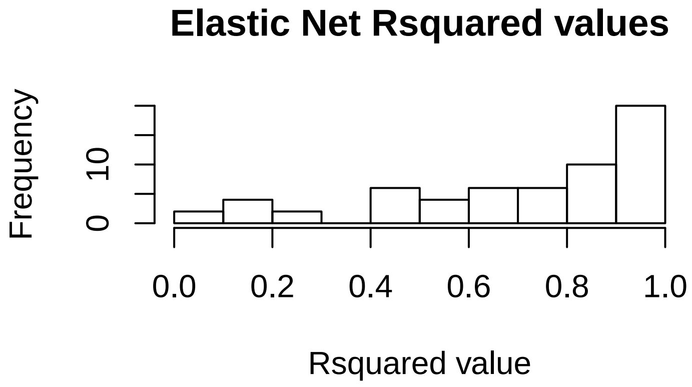




We see that theses results aren't very satisfactory, they are sometime less convincing that our baseline, and especially spread. Overall, the result isn't robust at all.

K-fold Method





Rsquared mean value:

```
## [1] 0.7025592
```

Here we see that the elastic-net results are better that the ones provided by the KNN method, with a Rsquared around 0.65 and a RMSE slightly under 3.

We also see that the Rsquared is squashed by very low Rsquared-values from seemingly badly made training/testing/validating set.

Conclusion

The Elastic Net and K-fold crossvalidation seem to provide robust results. They notably insure that the data and the target value are indeed related, and that the built model is capable of predict the tendancies of the traffic with the observation provided. Otherwise, it seems like some data set subsetting provide unsatisfactory results, which might be corrected via the exclusion of observation points or the acquisition of new data.

Alltogether, we assume that the best modelisation choice is to use the Elastic-Net algorithm as it shows the best results, and seems to be the most robust.