DREAM_CANCER_REPORT

TEAM

April 2015

1 Introduction

2 Data preparation

We had to deal with very messy data because of the presence of NA's at high rate and also missing values.

Variable	Pourcentage_NA
CREACLCA	96.75
SMOKSTAT	96.125
CREACL	94.0625
SMOKFREQ	92.8125
TSTAG_DX	72.3125
.....	.....
.....	.....
GLU	30.6875
REGION_C	30.125
NA.	30.0625
TBILI	1.4375
NEU	1.3125
TRT3_ID	1.1875
.....	.....
.....	.....
ALT	0.3125
CREAT	0.1875
WEIGHTBL	0.125
WGTBLCAT	0.125
ECOG_C	0.0625

The first Work was to complete and clean the data in a way to keep the useful information. We choose to keep SMOKE since this variable is a often kept in studies on other types of cancers.

3 Benchmark Model

On the clean data we apply a first benchmark model wich is the Cox model:

4 Feature selection

Goal :

- Find a sparse representation of the model

- Find new usefull variables
- Express qualitative data in numeric ones in order to use the know model

All along this work we use a criteria to determine the effectiveness of the construction

5 Predictive model for goal 1a

5.1 Models considered

5.2 Score and comparison

6 Predictive model for goal 1b

6.1 Models considered

6.2 Score and comparison

7 Conclusion