ENSAE ParisTech
Machine learning in computational biology
Professor : J.-P. Vert

Prostate Cancer Challenge report

Moussab Djerrab[1] **Lison Grappin**[2] **Jean-Benoît Emeyoud**[3] **Jérémy L'Hour**[4]

Contents

[1] moussab.djerab@ensae-paristech.fr
[2] lison.grappin@ensae-paristech.fr
[3] jean.benoit.eymeoud@ensae.fr
[4] jeremy.l.hour@ensae-paristech.fr

I Introduction

Prostate is a gland of the male reproductive system that plays a fundamental role in the production of sperm. Prostate cancer corresponds to the emergence of cancer cells in this gland. At the beginning, cancer cells mostly spread in tiny areas around the prostate. But if the medical treatment is not done early enough, cancer cells can migrate to other parts of the male body through blood vessels, and especially to lymph nodes. A metastatic castrate resistant prostate cancer (mCRPC) is declared when cancer cells have spread to parts of the body other than the prostate, and are able to grow and spread even though drugs or other treatments[5] to lower the amount of male sex hormones are being used to manage the cancer. Thus, mCRPC could be defined as an "advanced stage" (stage III or IV) of the prostate cancer with heavy complications since classical treatments have no effect on the propagation of cancer cells.

Though some treatments of mCRPC exist, including chemotherapies and supportive care, it remains unclear to determine the most effective single therapy or sequence of therapies. Their impact is modest with no improvement of overall or cause-specific mortality of mCRPC patients in the past 20 years. Moreover, prostate cancer is the most common cancer among men after lung cancer and colorectal cancer : for example, over the last ten years, 2 million men have been diagnosed in the USA, among which 15% had metastatic disease at the time of diagnosis. Among those 15%, approximately a third was concerned by mCRPC.

Recent research has promoted docetaxel[6] as a possible new treatment for mCRPC. Nonetheless, its real impact on the limitation of cancer cells propagation and on the overall improvement of mCRPC patients' health has not been formally proven yet. This is why the "Prostate Cancer DREAM Challenge" was launched by several American universities and Cancer research labs, among which the Prostate Cancer Foundation, to better analyze and understand docetaxel efficiency. The primary goal of this challenge is to provide a new prediction model to improve the inference on patients' survival according to the treatments they received, among which are docetaxel. We therefore have access to a database containing biological data of four phase III clinical trials with over 2000 mCRPC patients treated with first-line docetaxel. We chose to work on sub-challenges 1a and 1b concerning the probability of death and the time of survival of mCRPC patients.

Therefore, we first had to deal with the data, understand its contents and transform it to obtain a usable database to fit our models (II. Data preparation). Secondly, we fitted a benchmark model usually used in cancer survival prediction, namely a Cox-Lasso model, provided by Tibshirani (1997) and Halabi *et al.* (2014) (III. Benchmark model). Thirdly, we selected the best features that explained death probability and survival duration (IV. Feature selection). Finally, we tested and selected some new models to predict death probability (V. Predictive model for goal 1a) and survival duration (VI. Predictive model for goal 1b) of mCRPC patients.

[5]The most classical treatment used is called "Androgen Deprivation Therapy" (ADT).
[6]Docetaxel is a substance with anti-cancer properties derived from if's leaves.

II Data preparation

II.1 Presentation of the data

The data available to solve this challenge is essentially composed of a "Core Table" [7] on which we will fit our models to train them. A test database is also provided to test our models and submit our results to the platform. Four cancer trials of first line metastatic Hormone Refractory Prostate Cancer (HRPC) patients were used to create the dataset : 1 600 patients were followed during four years and three months and received docetaxel treatment in the comparator arm. These four sets of raw trial data were consolidated into the dataset called "Core Table", where clinically important covariates are captured.

Explained variables

Challenge 1a focuses on death probability in an horizon of x months. The dataset contains a dummy called Death indicating whether the patient died during the study (= 1) or survived (= 0). We can see on Table 1 that among the 1600 patients followed during the clinical trials, approximately only 41.44% (663 patients) survived after the final date observation of the clinical studies, which is relatively low. But this phenomenon can be understood as mCRPC is a final stage offering low opportunities to recover as classical drugs and treatments have no effect on the remission of cancer cells.

	Death	Survival
Number of patients	937	663
Percentage of the population	58.56%	41.44%

Table 1: Description of "DEATH" variable

Challenge 1b focuses on the survival duration of patients from the launching of the clinical trials, based on the variable called "LKADT_P". To study it, we calibrated Kaplan-Meier estimate (see Figure 1) in order to get the overall empirical survival function of patients targeted by the study. We can see for example that the probability to survive more than 4 years and 3 months (which corresponds to the duration of the study) is of approximately 20%.

Explanatory variables

The explanatory variables can be divided in several groups :

- Individual variables : age, race, Body Mass Index, height, weight, region of origin
- Behavioral variables : smoking frequency, degree of patient self-care
- Biological variables : rate of several antibody in blood (especially some that fight propagation of cancer cells), volume of glucose, calcium, sodium, magnesium, lymphocytes, etc. in the blood, blood pressure, etc.
- History of medication and drugs usage : dummies indicating prior use of radiotherapy, analgesics, corticosteroids ...
- Propagation of cancer cells in the patient body : dummies indicating whether the part of the body designated was infested by cancer cells or not
- History of medical problems and diseases : dummies indicating whether the patient ever had cardiac disorders, peptic ulcer disease, gastrointestinal bleed, etc. and dummies indicating genetic and family antecedents

[7] We chose to work on this table only because we considered that we do not have enough medical knowledge to include new biological variables that potentially could improve our models fitting.

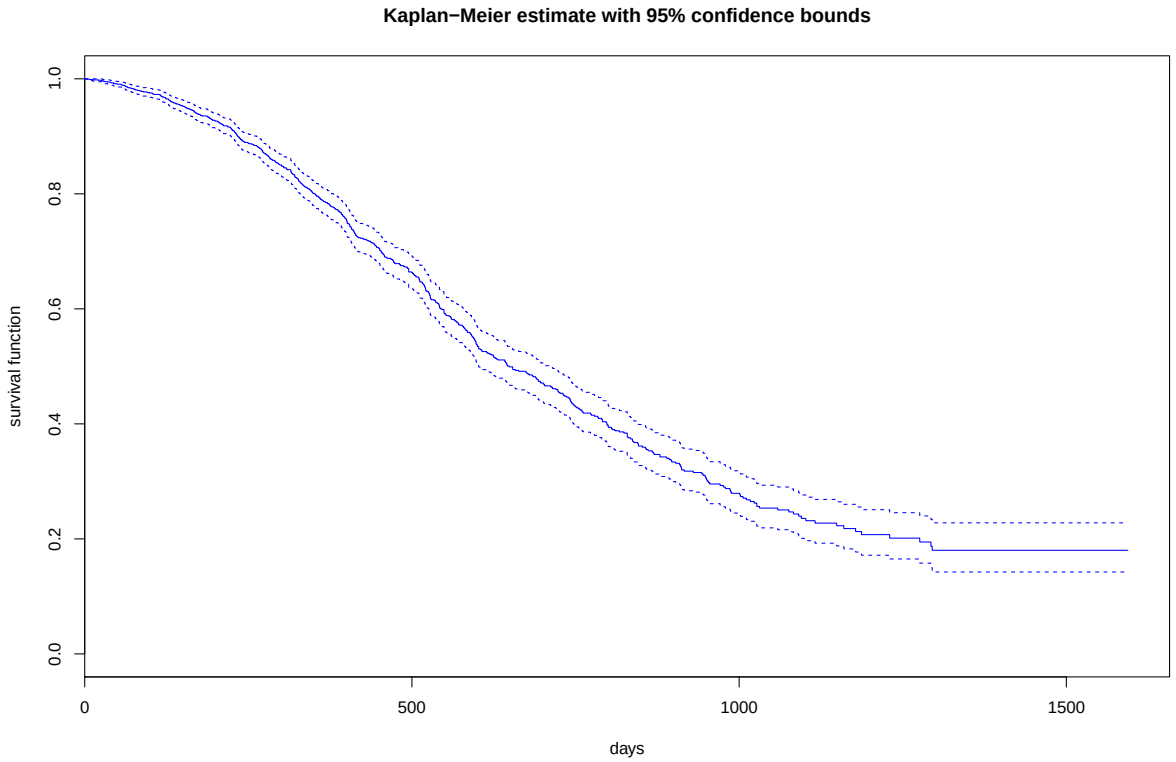


Figure 1: Kaplan-Meier estimators on survival duration

All these variables (more than a hundred) could potentially have an influence on death probability and survival duration. But variables are numerous, we will study in further parts ways to get both an efficient and parsimonious model to explain prostate cancer mortality.

First insights on possible links between explained and explanatory variables

In order to have some insights on the potential influence of explanatory variables on explained variable "DEATH" [8], we box-plotted some quantitative variables according to the value of "DEATH" dummy, and calculated some contingent tables crossing qualitative variables with "DEATH" dummy. Here are some results that we considered as interesting, and that helped us to better understand our sophisticated statistical models.

Among the quantitative variables, we found that the biological variables AST, LDH, TESTO, TBILI, WBC, CREACL, PHOS, RBC, LYM, CCRC, GLU, CREACLC, PSA [9] could graphically possibly have a an impact on the dummy "DEATH". An example of the boxplots realized is shown in Figure 2.

Regarding qualitative variables (especially individuals and behavioral variables) or dummies (such as history of medication, location of metastases cells ...), we computed contingent tables to infer the possible influences of those variables on "DEATH". For example, we studied the possible impact of variable "RACE_C" or variable "CORTICOSTEROID" on the dummy "DEATH". Here are the results, presenting the number of patients by cross-set and the column total (Example reading note : Among the Asian patients, 34.1% survived.). What is striking for the race variable for instance is that Asian patients seem to die more than other races : genetics could perhaps have an influence on the propagation of cancer cells in the body. At the same time, prior medical problems could have an impact on the probability of death as we can see in the case of prior corticosteroid.Nonetheless, these are only graphical intuitions, that must be infirmed or confirmed by the models fitted in the next parts of our report.

[8] We did the same type of work for "LKADT_P".

[9] In order to know the precise definition of each variable, please look at the dictionary of variables displayed on the synapse website.

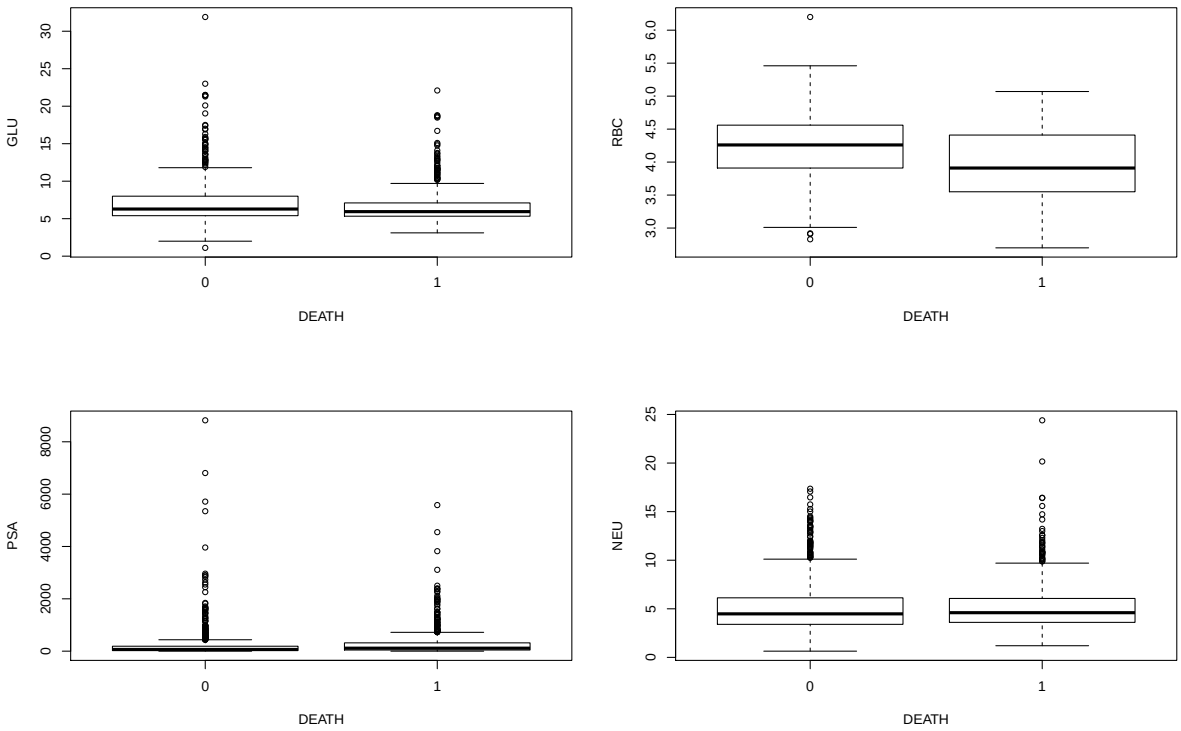


Figure 2: Boxplots to cross quantitative explanatory variables and explained variable "DEATH"

DEATH/RACE C	Asian	Black	Hispanic	Missing	Other	White
Survival	14	49	10	45	21	798
	34.1%	66.2%	71.4%	81.8%	80.8%	57.8%
Death	27	25	4	10	5	592
	65.9%	33.8%	28.6%	18.6%	19.2%	42.6%

Table 2: Cross-table of variable "DEATH" by the race of the patient

DEATH/CORTICOSTEROID	No	Yes
Survival	875	62
	62.5 %	31.0%
Death	525	138
	37.5%	69.0%

Table 3: Cross-table of variable "DEATH" by the dummy "Prior Corticosteroid"

II.2 Completion of missing quantitative data

As the dataset provided is a summary of 4 sources of data collected according to various methods, we had to deal with a lot of missing values. Table 4 presents the range of percentage of missing values for some explanatory variable. [10]

In order to improve the predictive power of our models, we chose to complete variables with missing a percentage of missing values below 50%, considering that completing a variable with more than a half of missing values could be detrimental to the model and create great bias during learning phase. We thus ignored variables with more than 50% missing values in our analyses except the variable SMOKE, because it is a benchmark variable used in classical cancer analyses.

[10] We do not explain exhaustively this data cleaning step of our work because it would be tedious and we preferred to focus on model fitting in this report.

We completed first numerical individual and biological variables among which are for example BMI, HEIGHTBL, WEIGHTBL, ALP, AST, ALT, CREAT, HB, LDH, TESTO ... Some of them had only less than 2% of missing data. We considered that the best way to complete those variables was to replace missing values by the mean of the variable over non-missing values. Other biological variables, that have approximately between 30% and 50% of missing data [11], we proceeded according the following steps :

- Analyze the predictive power of the variable to complete on DEATH and LKADT_P before completion, by estimating a logistic regression on DEATH and computing a linear regression and correlation coefficient for LKADT_P
- Test method 1 : Test completion by the mean over non-missing data and compute predictive power of the variable completed with the same technique as the previous step
- Test method 2 : Test completion using a linear regression of complete biological variables on the variable to complete and compute the predictive power as previously
- Select the best completion technique according to several criteria : minimal distorsion of predictive power calibrated between non-completed and completed variable, relevancy of the linear regression computed in method 2.

For example, we completed LDH by method 2 because it did not distorted relations between LDH and DEATH or LKADT_P as coefficients were around 0, and because R^2 coefficient of the regression used in method 2 was not too low.

Variable name	Percentage of missing values (in %)
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

Table 4: Check of the presence of missing values in raw data

II.3 Qualitative variables encoding and completion

Fisrt, we binarised all dummies (encoded "Yes"/"No" in raw data) in order to be able to include them in SVM and Lasso regressions easily. Second, we chose to complete only the qualitative variable SMOKE (variable to qualify the behavior of the patient toward smoking) even if more than 90% of the data is missing (because only one of the clinical studies reported it). We made this choice because we considered that this variable was fundamental to analyse prostate cancer propagation potential incentives, as it is

[11] Namely : LDH, TESTO, SODIUM("NA."), MG, PHOS, ALB, TPRO, CCRC, GLU.

classically used in other cancer medical analyses. We completed it using a logisitc regression of numerical biological variables over the variable SMOKE to complete.

III Benchmark Model

In this section, we display result from a simple Lasso Cox model where the features enters linearly in the log-hazard rate, as described in Tibshirani (1997). It is the model used in the benchmark article given by the organizers of the Challenge (Halabi *et al.*, 2014).

III.1 Lasso Cox model

The model used in Halabi *et al.* (2014) is the Cox proportional hazard model of Tibshirani (1997), where the conditional hazard rate is assumed to be of the form:

$$\lambda(t|x) = \lambda_0(t) \exp\left(x_i^T \beta\right) \tag{1}$$

$\lambda_0(t)$ is the baseline hazard rate that holds for all individuals. For individual i, the hazard rate is modified by a factor of $\exp\left(x_i^T \beta\right)$ according its own characteristics and coefficients of the model β. This parameter is usually estimated through maximization of the partial likelihood:

$$L(\beta) = \prod_{t \in D} \frac{\exp\left(x_{i_t}^T \beta\right)}{\prod_{i \in R_t} \exp\left(x_i^T \beta\right)} \tag{2}$$

where D is the set of indices of failure times, R_t is the set of indices of the individuals at risk (the survivors) at time t and i_t is the index of the individual that failed at time t. The Lasso Cox estimator maximizes this function while penalizing the ℓ_1 norm of the coefficient β so as to end up with a sparse solution:

$$\hat{\beta}^{CL} \in \underset{\beta}{\operatorname{argmin}} \; -\log\left(L(\beta)\right) + \lambda \sum_{j=1}^{p} |\beta_j| \tag{3}$$

λ is a scalar setting the degree of sparsity of the model: the larger it is, the sparser the obtained solution. Optimal choice of the penalty level has been studies extensively in the litterature but results are often of no use for the applied statistician. A practical method to set it is to use cross-validation. Amongst the assumptions of the model is that the censoring is non-informative, *i.e.* the censoring process is independent to the one that governs lifetime.

III.2 Estimation results

We display results from a Cox Lasso model on the dataset separated into a training part and a test part.

Firstly, we run a Cox Lasso on the training dataset to select the covariates to enter in the hazard rate amongst the 103 variables. In order to set the penalty parameter, we used a 5-fold cross validation that ended up selecting a model with 18 non-zero coefficients. In order to remove the bias from the penalty term, we used a Post-Lasso estimate. In other words, we ran a simple Cox model using only the 18 variables previously selected. Results are displayed in Table 5.

As it is often the case when the Lasso is used as a selector, not all selected variables are statistically significant: only 14 of the 18 variables are significant at the 5% level or below. The R^2 is relatively low regarding the fact that we are concerned with prediction of the dependent variable. On the plus side, it is also a sign that we are not over-learning the training dataset. Regarding the interpretation of some coefficients, we find that smoking increases the risk of a fatal outcome by 27%, past history of cerebrovascular accident increases the risk by more than 100%, having liver and adrenal lesions also increases the hazard rate. Prior us of Gomadotropin and the sodium rate are associated with lower risks. Other coefficients are more difficult to interpret without medical knowledge, such as the cofficients on the presence of Congenital Disorders (lowers the hzarad rate!), history of vascular disorders, history of endocrin disorders just to cite a few.

Table 5: Post-Lasso Cox Model

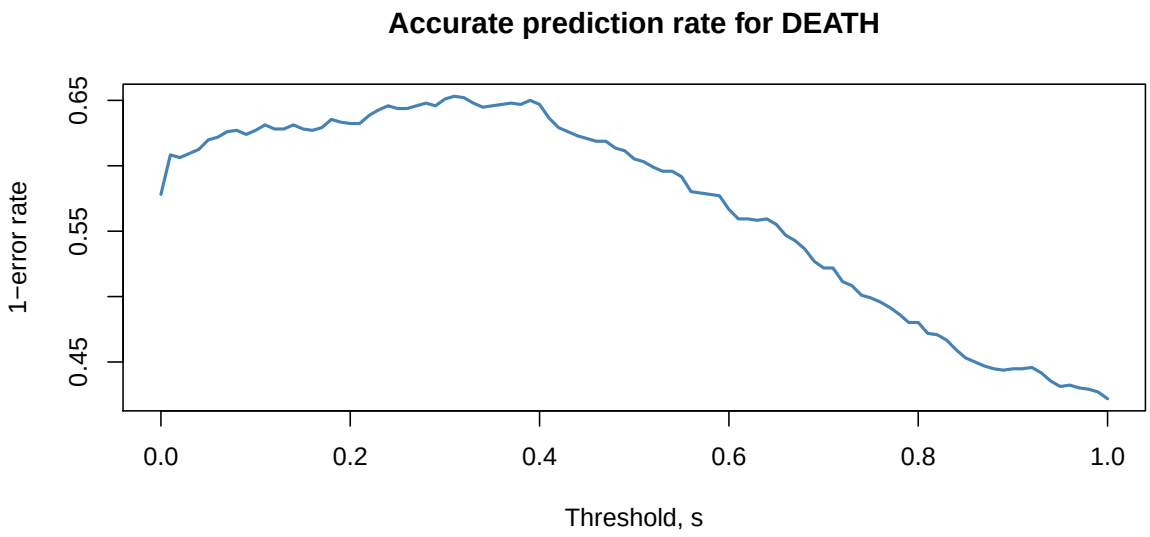
	Dependent variable:
	LKADT_P
CEREBACC	0.777***
	(0.264)
SODIUM	−0.020
	(0.020)
MHENDO	−0.738**
	(0.313)
ENTRT_PC	−0.006***
	(0.001)
HB	−0.196***
	(0.035)
GONADOTROPIN	−0.455***
	(0.142)
MHVASC	−0.235**
	(0.106)
LIVER	0.465***
	(0.171)
ECOG_C	0.240***
	(0.092)
PSA	0.0001
	(0.0001)
ESTROGENS	−0.394**
	(0.168)
ALB	−0.021
	(0.015)
LYMPH_NODES	0.279***
	(0.102)
LDH	0.001***
	(0.0002)
ADRENAL	0.787**
	(0.312)
ALP	0.0003***
	(0.0001)
SMOKE	0.245**
	(0.109)
MHCONGEN	−0.698
	(0.510)
Observations	960
R^2	0.306
Max. Possible R^2	0.993
Log Likelihood	−2,219.962
Wald Test	341.200*** (df = 18)
LR Test	350.579*** (df = 18)
Score (Logrank) Test	347.339*** (df = 18)
Note:	*p<0.1; **p<0.05; ***p<0.01

III.3 Prediction performance

Prediction of death In this part, we are interested in predicting the death of each individual. It is a rather complicated task. The first reason is that some observations are censored so when we don't observe the full survival duration of an individual it may be because he survived because he beat cancer or because the study stopped before we could witness its death due to cancer. The Cox model accounted for that fact. The second reason is that the Cox model targets the hazard rate of the individual, not directly the binary event of death or the total expectation survival duration. Hence we have to take the estimated hazard rates for different periods of time and make something out of them to predict the death variable.

Since the challenge organizers want as an output predicted death probability at 12, 18 and 24 months we will compute them based on the model estimated in Table 5 and use them to figure out whether we think it is more likely that a given individual has died. Our method is the following. We take a given threshold in the interval $(0, 1)$ that we call s. We predict death for an individual if and only if at least one survival probability amongst the three (12, 18 and 24 month) is below this threshold. To calibrate this parameter s, we minimize the error rate (obtained using the 0-1 loss function) on the training sample. Figure 3 displays 1 - the error rate as the threshold varies. The maximum is found to be for a threshold of .31. Table 6 displays the performance of the model and the training and test datasets. On the training dataset, the model is right about two third of the time, while this performance drops significantly for the test dataset where the model is right only about 40% of the time.

Figure 3: Accurate prediction rate for DEATH

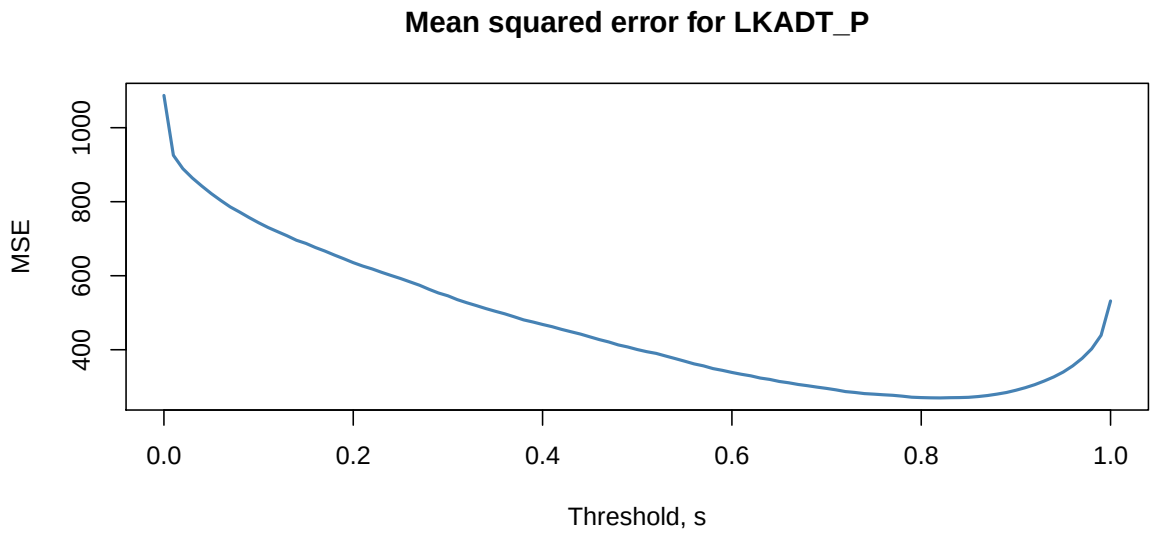


Threshold	.5	.31
Training dataset	0.61	0.65
Test dataset	0.38	0.40

Table 6: Accurate prediction rate

Prediction of duration For the prediction of the expected duration, we take a similar approach: we compute all the survival probabilities at each day between 0 and 4.1 years for every individual in the sample. Then we estimate the duration as the rank of the first day for which the probability of surviving falls below a given threshold s. The threshold is calibrated by minimizing the Mean Squared Error on the training sample as displayed in Figure 4. The MSE is minimized for a value of the threshold which is .82. It means that if the survival probability of an individual drops below .82 on a given day, we assume that the individual dies this day. The performance of the model on training and test datasets is reported in Table 7. Figure 5 displays the prediction of the model compared to the actual duration at the optimized threshold. The red line is the 45 degree line.

Figure 4: MSE for LKADT_P



Threshold	.5	.82
Training dataset	400.57	269.8
Test dataset	414.72	286.4

Table 7: MSE for prediction of duration

Figure 5: Prediction LKADT_P: training

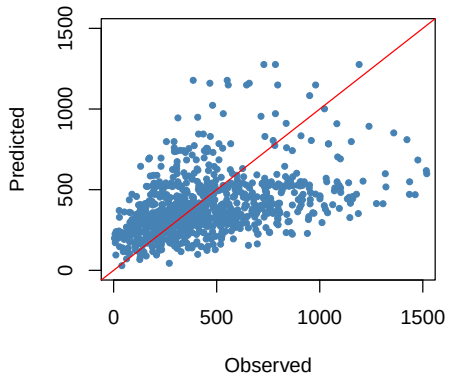
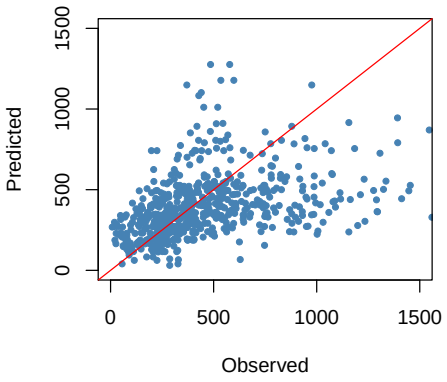


Figure 6: Prediction LKADT_P: test



IV Feature work

Goal :

- Find a sparse representation of the model
- Find new useful 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 which is the score of the best model (Random classifier on the target "DEATH").

IV.1 Dealing with Qualitative Data

In the data set we found mainly numerical data but a few are qualitative. In the tradition of statistics and econometrics is to put a random order into the data . In our work we test this approach and another one which consists in replacing the data by a numerical one based on the corresponding frequency of other numerical features.

Projection of the qualitative data into the numerical variables space, by the way of mean

Numerical data

i	Num_1	Num_2
1	45	75
2	78	33
3	25	38
4	10	10

Mean vector per unique qualitative element

Qualitative data

i	$VarQual$
1	A
2	B
3	A
4	C

ACK

i	$VarQual_{mean(Num_1)}$	$VarQual_{mean(Num_2)}$	Num_1	Num_2
1	35	56,5	45	75
2	78	33	78	33
3	35	56,5	25	38
4	10	10	10	10

By this technique we are able to add qualitative information to predictive model without adding random order. Also we choose to apply a mean function in order to introduce features of the same numerical scale than the others so as not to have the new information be crushed. The drawback of this method is that we make the size of the matrix increase very fast (according to the number of unique elements in each qualitative feature). In our setting we increase the size of the matrix by multiple factor of 10.

IV.2 Shannon reduction

High dimension of feature space is often a problem in machine learning task. Here we try an approach of reduction based on a trade off of performance and information reduction. First we take the "DEATH" target in order to be the subject of information. We thus construct densities out of the other variables :

$DEATH$	Var_1	Var_2	...
0	p(DEATH = 1 — Var_1)	p(DEATH = 1 — Var_2)	...
1	p(DEATH = 0 — Var_1)	p(DEATH = 0 — Var_2)	...

In order to score the variables we measure the Shanon entropy for each variable:

$$-\sum_{i=1}^{n} p_i log(p_i)$$

Then we keep the top 10% which means in our keys almost 100 variables out of the 1000 that we created. This enables us to sparsify the model the result for both goals (1a and 1b)

IV.3 Features importance reduction for RandomForest

Using the Random Forest technique we can extract the feature importance for tasks:

Variables	Feature$_I mportance$
LDH	0.21
PER_REF	0.06
ALP	0.04
$LKADT\_REF$	0.04
PSA	0.03
$ENTRT\_PC$	0.03
HB	0.03
$WEIGHTBL$	0.02
BMI	0.02
NEU	0.02

This table shows the top 10 of the variables that have the more effect on the DEATH of the patient. In the next part we are going to present the predictive models that we have used

V Predction implementation and scoring

V.1 Models considered for goal 1a

Goal 1a is about predicting the death of a patient. Therefore we can achieve this task by implementing some classifier learners. Because of the structure of the data we'll use only the numerical data First. Then we'll show the result obtained when adding the qualitative data. Several models have been tested among which : Logistic regression, SVM classifier with different kernels,Random Forest Classifier.

V.2 Score for goal 1a

Model	Score
SVM classifier (rbf Kernel)	0.0575
SVM classifier (poly Kernel)	0.29573
SVM classifier (Linear Kernel)	0.4146
Logistic regression	0.7578
Random Forest Classifier	0.865

These results show that a good separation for the classification task should be linear. Among the models two seem to stand out : the logistic regression and the Random Forest classifier. The last one enables us to extract the feature importance for predictiting death. Whereas the Logistic regression enables us to see the sign of the effect of each of those wariables.

V.3 Models considered for goal 1b

Here, as said before, we 'd like to predict the survival duration. In order to do so we apply two different model which are :

- The cox model
- Classical models of regression

Also we'll consider the l2 norm as loss functions. In the following table we show the result for each method:

Model	Score
Linear regression	0.49
SVM regressor (linear Kernel)	0.41
Random Forest Regression	0.68
Cox model	0.30

VI Conclusion

Several Methods have been used in order to obtain small real improvement of performance in the end. We had not the time to test the individual data , but this would have required structural prediction. Each patient can be represented as a graph each node being one of his characteristics, the edges being weighted by the frequency of link. Under this framework we'll need to define a kernel for graphs. In the remaining time of the competition we'll try this approach.

Also our code is a mix of R scripts (for the stats and the econometrics) and python scripts (for the machine learning part). All the source code as well as the report latex and pdf is in open source on the repo git :

https://github.com/MoussabDjerrab/Projet_DREAM_CANCER.git

VII Bibliography

Halabi, S., Lin, C.-Y., Kelly, W. K., Fizazi, K. S., Moul, J. W., Kaplan, E. B., Morris, M. J., and Small, E. J. (2014): "Updated Prognostic Model for Predicting Overall Survival in First-Line Chemotherapy for Patients With Metastatic Castration-Resistant Prostate Cancer". *Journal of Clinical Oncology*.

Tibshirani, R. (1997): "The Lasso Method for Variable Selection in the Cox Model". *Statistics in Medicine*, 16(4):385–395.