INRA Predict Workshop

TCGA (The Cancer Genome Atlas) Predictive modeling using small data

MyDataModels - Sophia Antipolis- France
www.mydatamodels.com

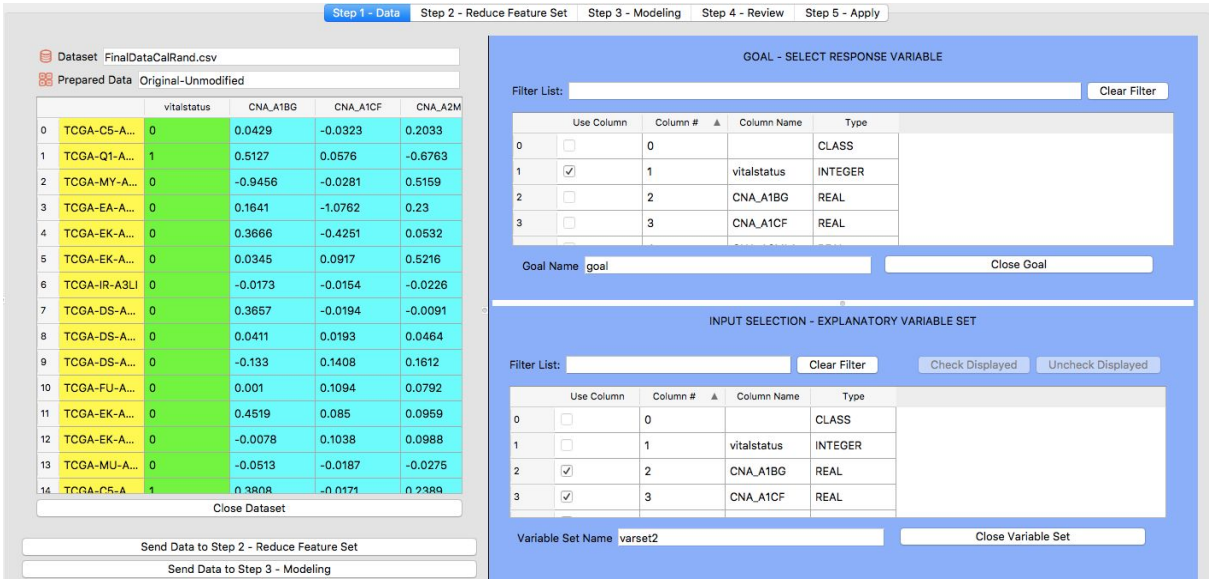
MyDataModels participated to an INRA (Institut National de la Recherche Agronomique) PREDICT workshop where the provided dataset was an extract of the TCGA (The Cancer Genome Atlas) database. Dataset addressed an uterus cancer case. It contained four groups of variables, with no missing values.

Objective of this workshop was to demonstrate that MyDataModels automated Machine Learning solution could be used first to find the top 25 variables impacting resulting variable, and second to produce a high accuracy predictive model.

The entire process took 3 days on a MacBook Air and was run by a person having no coding and no machine learning skills.

First the provided CSV file was imported in the MDM workspace.
Dataset size was 64Mb, contained 90 samples - women - with 53,000+ genomic variables. The entire process was run on a standard MacBook Air.

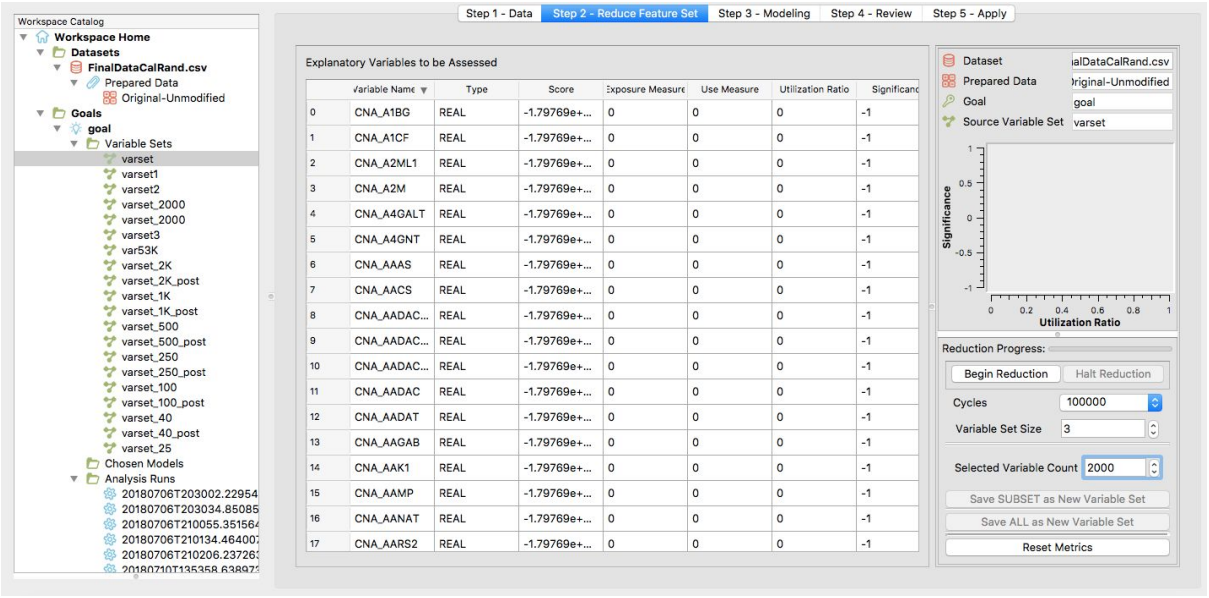
Next, a Goal was set specifying the VitalStatus (0=negative uterus cancer, 1=positive uterus cancer) variable as the analysis objective. 21 positive cases (23%), and 69 negative cases (77%).



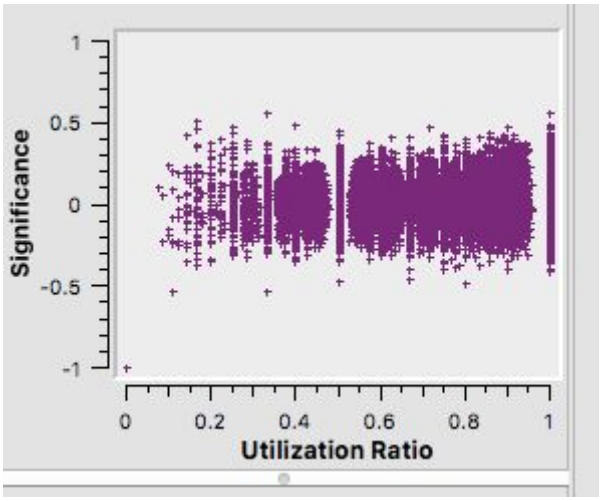
Feature reduction process before modeling

Feature reduction process described hereafter took 3 days.

The Variable Set Size parameter was set to a value of “3”. The parameter Cycles was set to a value of 100,000. Feature reduction was then performed for approximately four successive iterations of 100,000 cycles.



At this point, the Significance/Utilization (S/U) scatter plot presented the following visualization which indicates good differentiation of the variables under consideration:

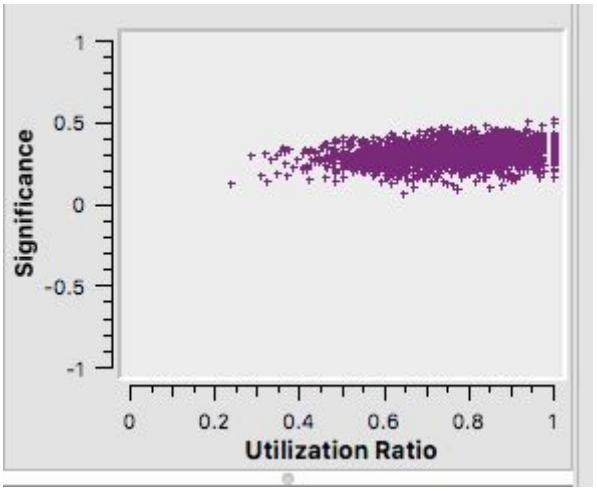


As there was good differentiation depicted in the S/U scatter plot, we saved the top 2000 variables identified in the reduction in a new variable set. This was done for performance reasons. As there was good differentiation observed, we can safely discard the bulk of the least significant variables. There is no value in spending computational resources determining the relative merit of the least significant variables. The goal was to identify and rank only the most important variables.

This new variable set of the top 2000 variables was then loaded for reduction in the Step 2 tab. The Reset Metrics button was used to discard the reduction metrics on this new variable set so that reduction could focus solely on the relative merits of these 2000 variables.

The Variable Set Size was again set to a value of “3” and cycles to a large number so that reduction would continue without intervention until manually halted. Reduction was run until the average Exposure Values listed in the variables table was approximately 25.

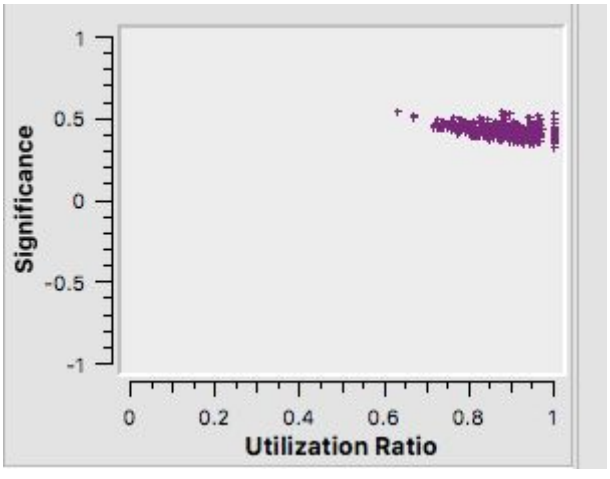
At this point the S/U scatter plot again indicated that the variables were well differentiated:



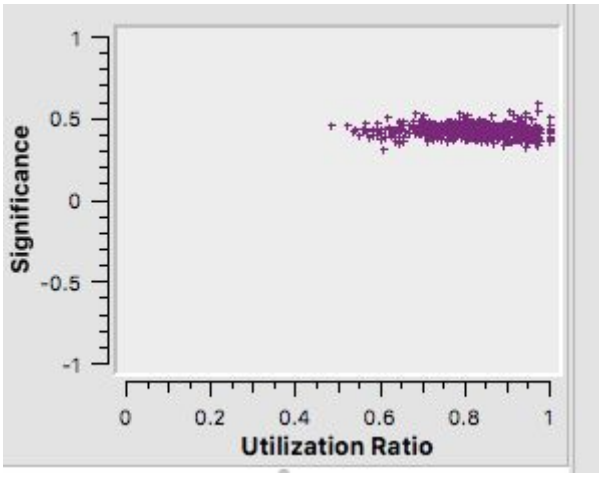
This iterative process was repeated to produce new variable sets of the top 1000, 500, 250, 100 and finally the top 25 variables.

Each iteration entailed the above described process: the selection of a subset from the previous iteration, loading that subset for reduction, resetting the metrics, setting variable set size to 3, running reduction cycles until good differentiation is observed in the S/U scatter plot. A good practice is to use a naming convention that indicates each step of these iterations. The S/U scatter plots for each of these stages as we performed them are shown below.

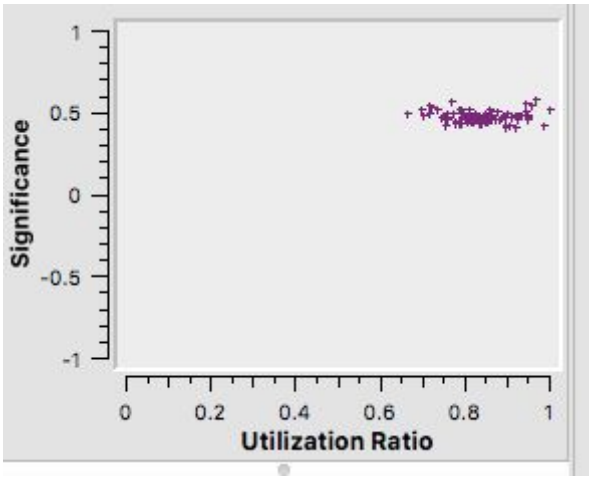
Top 1000:



Top 500:



Top 100:



Results in less than 1 hour

This step took less than 1 hour.

After final reduction on the top 100 variables, the top 25 variables were saved to a new variable set and sent to modeling in Step 3 tab.
Within 12 attempts, we decided to select the best classifier versus what we thought was the performance criteria, True Positive Rate. It was found to predict VitalStatus (Model ID: 20180716T173718.030587) with 74% accuracy on an hold out subset, but more importantly with True Positive Rate (Sensitivity) at 83%, 1 error on 6 samples.

Performance Metrics

	MCC	ACC	TPR	TNR	FPR	FNR	PPV	NPV	F1	P	N	TP	TN	FP	FN	ideterminat	AUC
TRAINING	0.59485	0.805556	1	0.766667	0.233333	0	0.461538	1	0.631579	6	30	6	23	7	0	0	0.977778
VALIDATION	0.426401	0.666667	0.888889	0.555556	0.444444	0.111111	0.5	0.909091	0.64	9	18	8	10	8	1	0	0.783951
TEST	0.463348	0.740741	0.833333	0.714286	0.285714	0.166667	0.454545	0.9375	0.588235	6	21	5	15	6	1	0	0.904762
ALL DATA	0.511058	0.744444	0.904762	0.695652	0.304348	0.0952381	0.475	0.96	0.622951	21	69	19	48	21	2	0	0.900621

We can see in the above table accuracy obtained on test subset is 74% with True Positive rate at 83%, True Negative rate at 71%, and AUC at 90%.

We then scored the selected model on a 45 samples test dataset, 11 positive cases and 34 negative cases, where VitalStatus variable content was empty. Results obtained with selected model were compared to actual results and accuracy found was 67%, True Positive Rate at 91% (10 predicted correctly on 11) and True Negative Rate at 59%.
We believe we could further improve performance by using a stratified sampling method and ensemble learning.

Selected model used only the 4 following variables - out of 25 top variables - and took following compact functional form.

4 variables Used:

CNA_PDSS2
RNASEQ_FAM157A
MIRNA_hsa-mir-122
MIRNA_hsa-mir-518f

MODEL FUNCTIONS:

Each function computes a propensity score for each of the 2 cases, scores difference leads to prediction

OUTPUT1: *EXP('CNA_PDSS2') - 0.4680247196154727 * (EXP('CNA_PDSS2') - ('CNA_PDSS2' - SIN('MIRNA_hsa-mir-518f') - 0.08503852903028916 * (SIN('MIRNA_hsa-mir-518f') - INT('MIRNA_hsa-mir-518f'))))*
OUTPUT2: *ABS('MIRNA_hsa-mir-122') - 0.3956511787594415 * (ABS('MIRNA_hsa-mir-122') - COS('MIRNA_hsa-mir-122'))*

List of 25 most significant variables used to produce resulting model is below:

VARIABLE_NAME	***BINARYCLASS_SCORE***
CNA_WASF1	0.477964521
CNA_SLC16A10	0.476000284
MIRNA_hsa-mir-518f	0.472054263
CNA_C6orf203	0.471001985
CNA_PPIL6	0.469287161
RNASEQ_TLX1NB	0.467656627
CNA_PDSS2	0.464896603
MIRNA_hsa-mir-122	0.464621088
CNA_NR2E1	0.449025443
RNASEQ_FAM157A	0.449022432
MIRNA_hsa-mir-526b	0.44823409
METHYL_CNTN2	0.439968833
METHYL_CCNDBP1	0.439905797
METHYL_XIAP	0.428070469
CNA_NDUFB1	0.427942846
CNA_CPSF2	0.424530422
MIRNA_hsa-mir-517a	0.421886422
CNA_PSMC1	0.419892563
METHYL_RPL36A	0.414748999
METHYL_CBX2	0.408613885
MIRNA_hsa-mir-1269	0.406860894
METHYL_ZNF682	0.406782143
RNASEQ_NUBP1	0.406386016
METHYL_UBE2A	0.397948095
METHYL_AMMECR1	0.396135335

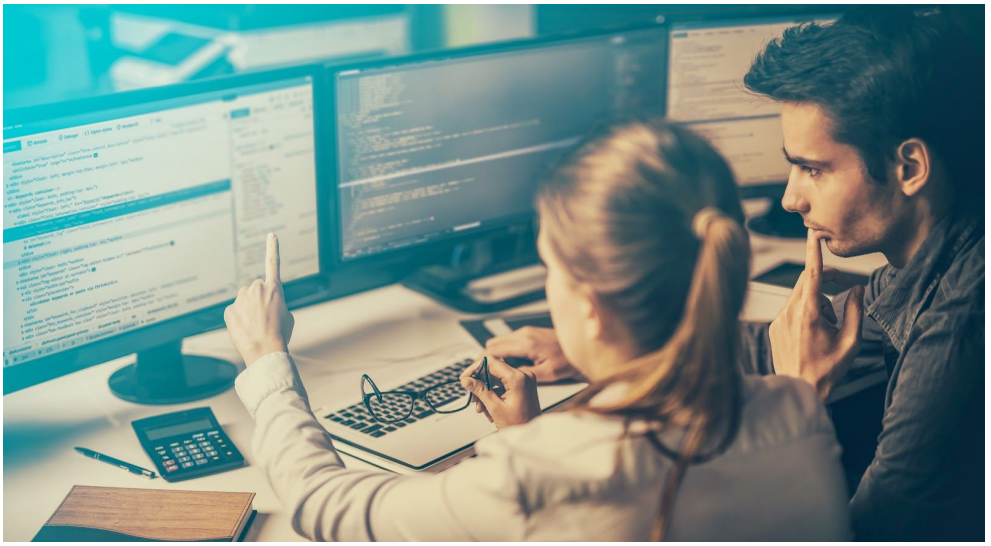
Summary

- Case study: model generation for uterus cancer prediction using historical data
- Historical dataset size: 64Mb, 90 samples, 53 161 variables
- Variables: 15 163 Methylation, 852 miARN, 17 130 CNA, 20 016 RNAseq
- Equipment used: MacBook Air, 1.8 Ghz Intel core i5, 8Gb 1600Mhz DDR3
- Variable reduction process: 3 days
- Modeling process: 1 hour
- Model's accuracy: 74%, with 83% true positive rate, 71% true negative rate

MYDATAMODELS

The Automated Machine Learning Company

MyDataModels is a software company that develops and markets an Automated Machine Learning 2.0 Platform.
Automated Machine Learning 1.0 opened the path few years ago. It automates most of the predictive modeling process, but still requires some coding and Machine Learning skills.
MyDataModels goes much further. Our Automated Machine Learning 2.0 service is powered by our unique Machine Learning engine, inspired by evolutionary algorithms. It is a one click data-in model-out service which enables all professionals to build predictive models without coding or Machine Learning skills.



Our Partners:

bpifrance Inria aws SOLUTIONS COMMUNICANTES SECURISEES Cluster IA