Geneva University Hospital blood cancer biomarker analysis and modeling case study

---

Predictive modeling
using small data

MyDataModels - Sophia Antipolis- France
www.mydatamodels.com

Mantle cell lymphoma (MCL) and chronic lymphocytic leukemia (CLL) share many features and both arise from CD5+ B-cells; their distinction is critical as MCL is a more aggressive neoplasm.

Thomas Matthes, Professor of Hematology at Geneva University Hospital, and Sierrabolics have teamed up to help Professor's research team in his gene analysis and root/cause analysis work related to MCL and CLL diseases.

The research team had assembled a very small dataset, which contained 50 samples (20 MCL and 30 CLL) with 290 attributes per sample. The attributes were obtained using Nanostring technology® and corresponded to values for gene expression levels. The dataset file size was 150 kilobytes.

Typically, producing a predictive model from a very small dataset (like the one used in this case) is extremely challenging via traditional machine learning techniques.

This is a common difficulty in such research due to the dilemma that common analysis techniques require substantial numbers of samples in order to attain useful analysis results yet the procurement of each sample can be both expensive and logistically difficult.

As such, the dataset under consideration offered an excellent test case to challenge the efficacy of MyDataModels under difficult analysis constraints.

MyDataModels has been used 1) to find the top 25 genes among the 290 responsible of the MCL and CLL diseases, and 2) to build a predictive model, which will allow clinicians to diagnose automatically MCL or CLL diseases on future patients for whom the gene expression levels of 290 genes have been determined.

Preparation of the dataset file

The response variable Diagnosis (MCL or CLL) value was derived from the metadata. Rows were then randomized such that the order of samples in the rows was ensured to be random.

Results in less than 1 hour

Feature reduction, (which in MyDataModels produces a ranking of the entire set of feature attributes in terms of each feature's ability to predict the response variable) and generation of an explanatory model, took just under an hour. When evaluated against an independent hold out set of 15 samples, the resulting explanatory model was 100% accurate with a true negative rate of 100%, and a true positive rate of 100%.

With regards to the ranked feature set, in the top 25 genes identified by the analysis process, the research team found that there was a 90% overlap between their previous manual work and MyDataModels results: in the 10% delta, some genes found by MyDataModels as top25 were not found previously by R&D team, and some top 25 genes found by R&D team were not found by Databolics. This could not be interpreted immediately as errors from one side or the other, it might happen that genes found by MyDataModels were actually in the top 25 ranking or the other way around. Further analysis is currently being conducted as well as increasing the 50 samples dataset with additional patients to get to more precise conclusions.

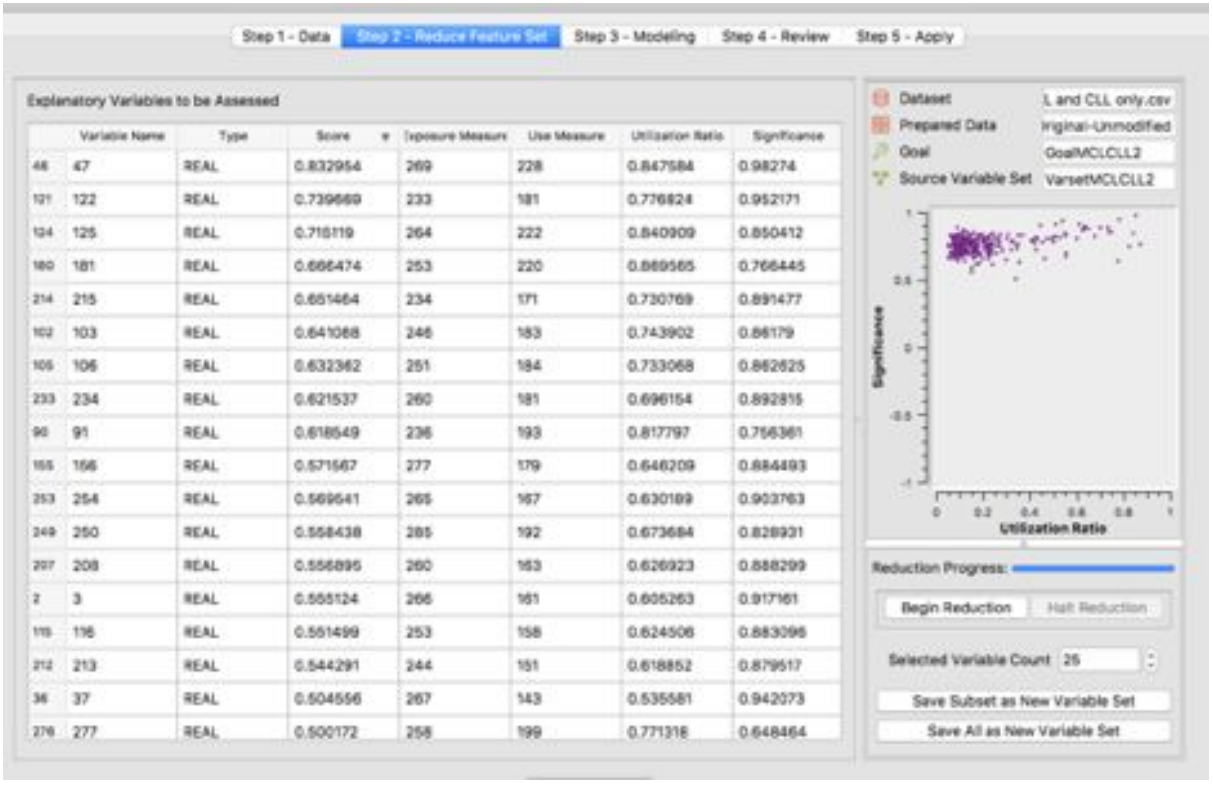


Figure 1: *results of the top 25 genes ranking process, highlighting the gene number by order of importance (from highest score)*

Professor Matthes's quote: *"Before using MyDataModels, we were using manual intensive and time consuming algorithmic methods to identify genes responsible of blood cancer pathologies. MyDataModels brought us speed, accuracy of predictions and ease of use. As clinicians and researchers, we have neither time nor programming and IT skills to use any of the predictive technologies available today. Having an automated predictive modeling tool makes a huge difference in our research work from timing and cost perspectives. We will keep on using MyDataModels to refine our models, and will apply it to other research works. We also plan in the future to make the predictive models available to clinicians to support their diagnosis processes."*

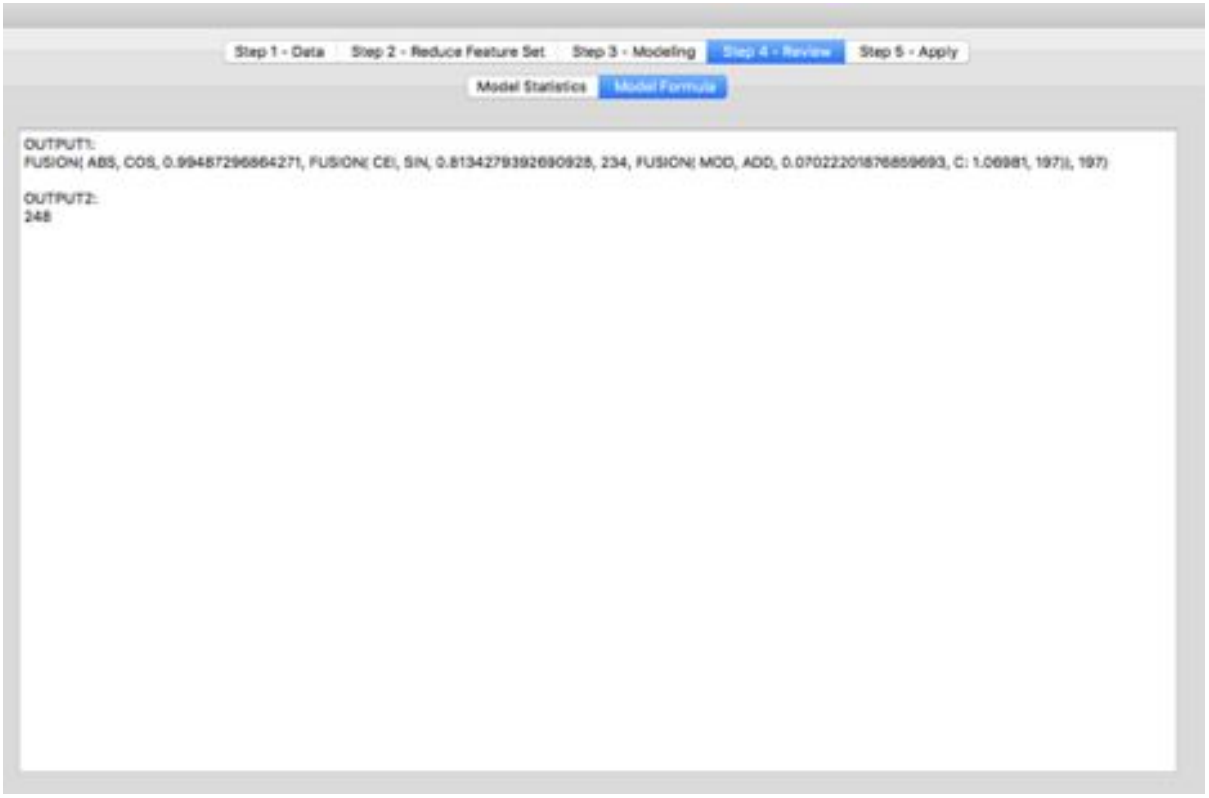


Figure 2: *resulting model. It uses 3 variables (genes "234", "197" and "248"), several operators and constants. In a binary classification model, model is made of 2 equations, one calculating probability of true prediction, the other calculating probability of false prediction. Upon computing the value of each equation, the greater value yields the case predicted by the model. The difference between these 2 numbers gives the prediction's confidence score.*

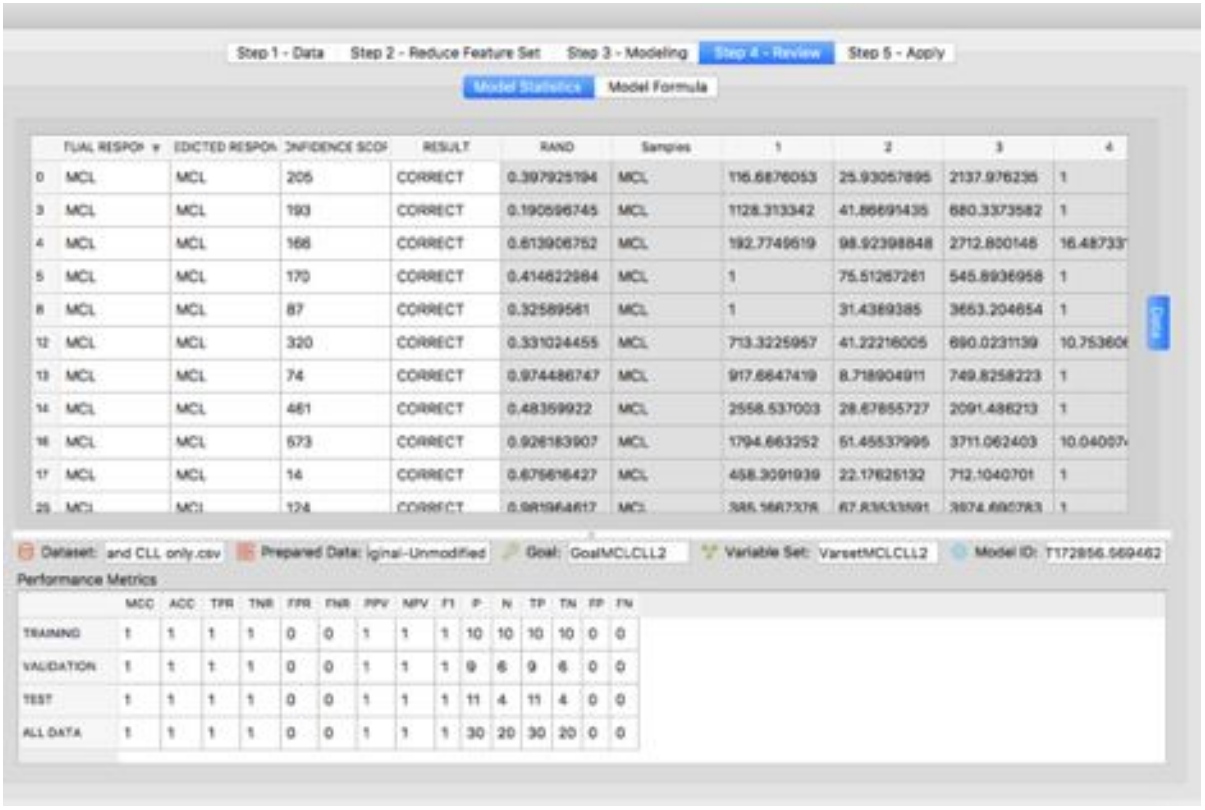
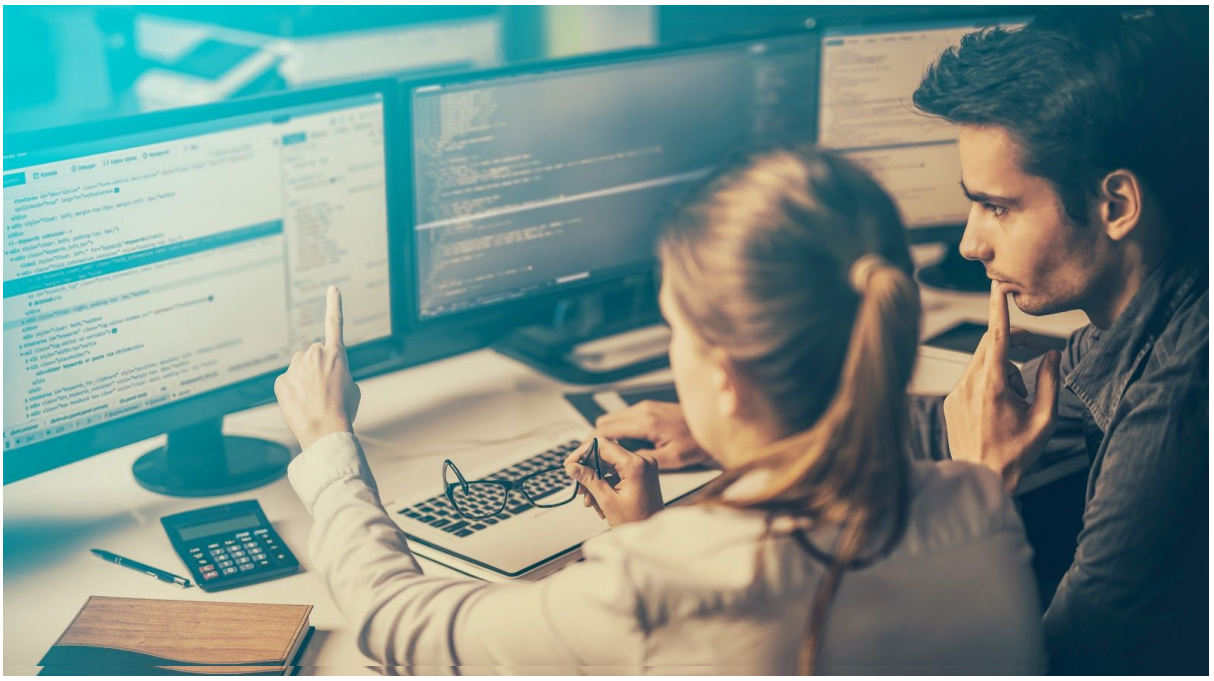


Figure 3: *Review of best model statistics. One can see the 100% accuracy obtained on the entire dataset and on the 3 subsets (training, validation and test, training and validation are used for model creation, the test set is an independent hold out used to provide a final accuracy assessment by the final model again new data not available to the modeling process) used to build hundreds of potential models and finally select the best model.*



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.

Cluster IA

SOLUTIONS COMMUNICANTES SECURISEES
PÔLE DE COMPETITIVITE MONDIAL