MYDATAMODELS

Genomics case study

Biomarkers discovery

MyDataModels - Sophia Antipolis- France
www.mydatamodels.com

MyDataModels was used to model and perform feature selection on the NCBI microarray gene expression dataset GSE19429 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE19429).

This dataset contains **Affymetrix GeneChip Human Genome** V133 Plus 2.0 microarray data representing gene expression levels for 200 samples. These samples consisted of bone marrow tissue obtained from 183 patients with myelodysplastic syndromes (MDS) and 17 healthy controls. The microarray used generated 54,675 attributes per sample. The dataset file size was 103 megabytes.

Acquisition and Preparation of the dataset file

This entailed download of the series matrix text file, removal of metadata rows, transposition of rows/columns such that columns represented attributes and rows represented samples. The response variable value was derived from the metadata and added as a column.

Rows were then randomized such that the order of samples in the rows was ensured to be random. The healthy vs MDS attribute was chosen as the response variable to analyze.

Results in less than 2 hours

This included dataset acquisition and preparation, feature reduction, ranking of the entire set of feature attributes and generation of an explanatory model, took just under two hours. When evaluated against hold out samples (samples not available for consideration in the analysis process), the resulting explanatory model was 90% accurate with a true negative rate of 100%, a true positive rate of 89.1%.

With regards to the ranked feature set, in the top 20 probes identified by the analysis process, 4 probes (231067_s_at, 241679_at, 210517_s_at, and 227530_at) were identified that represent transcripts for the gravin/AKAP21 gene, relevant to MDS as discovered by other research:

http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2141.2004.05067.x/epdf
and
http://www.nature.com/leu/journal/v24/n4/full/leu201031a.html .

This is of note because MyDataModels placed 4 of the 5 probe ID's for AKAP21 in the top 20, which lends confidence that the analysis process is truly considering the merit of all attributes and not simply randomly finding useful attributes.

Also in the top 20 attributes was found three probe IDs that represent transcripts for ARPP21 (220359_s_at, 1556599_s_at, and 231935_at). The relevance of ARPP21 to MDS was also found by the research described in:

http://www.nature.com/leu/journal/v24/n4/full/leu201031a.html .

Also identified by MyDataModels in the top 20 probes:
OR7A5 (208285_at) - A finding also found by:
http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2141.2007.06833.x/pdf

SH2D4B (1563849_at) and KIAA0226L (previously named C13orf18, probe 44790_s_at) - Both found to be down regulated and differentially expressed in MDS patients per:
http://bmcmedgenomics.biomedcentral.com/articles/10.1186/1755-8794-3-30

PPP2R2C (228010_at) - Downregulated in MDS patients per:
http://onlinelibrary.wiley.com/doi/10.1002/ijc.27896/full
CD19 (206398_s_at) - Found to be downregulated in MDS patients per:
http://ajcp.oxfordjournals.org/content/138/5/732
P4HA1 (202733_at) - Found to have gene pathway aberrantly methylated in MDS HSCs:
http://www.bloodjournal.org/content/bloodjournal/120/10/2076.full.pdf?sso-checked=1
TP53INP1 (225912_at) - Relevant mutation characteristics in MDS:
http://williams.medicine.wisc.edu/mdsgenetics.pdf
TP53 as prognostic biomarker and association with higher likelihood of transformation to AML:
https://www.mycancergenome.org/content/disease/myelodysplastic-syndromes/tp53/331/

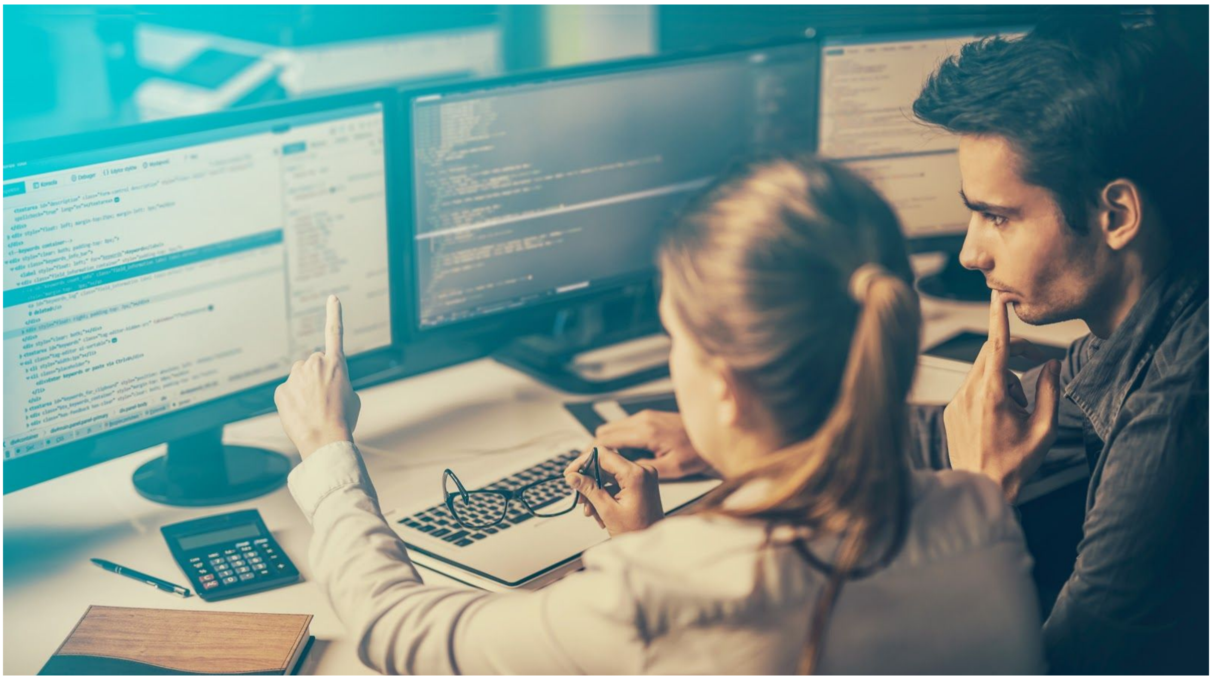
And

http://www.mdsbeacon.com/news/2014/05/08/p53-protein-levels-prognosis-lower-risk-mds-with-del-5q/[1]
IFR4 (204562_at) - Relevance to MDS per:
http://www.bloodjournal.org/content/124/21/2203.abstract?sso-checked=true

and

http://www.bloodjournal.org/content/122/21/1529.short?sso-checked=true

The following probes/genes were also identified in the top 20 probe IDs but relevance not found in other research or literature:
Probe: 1568611_at
Gene: HMHB1, Probe: 208302_at
Gene: DUSP26, Probe: 219144_at
Gene: MME, Probe: 203434_s_at
Gene: P2RY14 Probe: 206637_at



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