Predictive Risks of Colorectal Cancer by Machine Learning

Asia Pacific Electronic Health Records Conference

17-18 Oct 2019

John Mok

Health Informatics
(Standards & Policy 3)

IT&HI DIVISION

Acknowledgements

- Hong Kong Hospital Authority
  - Dr NT Cheung, Head and CMIO of IT&HI Division
  - Ms Vicky Fung, Senior Health Informatician
  - IT&HI colleagues

Outline

- Background
- Design
- Data science tools
  - Weka & DataRobot
- Results
- Lessons learnt

Background

- A Proof of Concept study was conducted last year – the objective was to gain some practices in Machine Learning with a clinical use case.

The RESULTS of this paper was our target

Dig Dis Sci. 2017 Oct;62(10):2719-2727. doi: 10.1007/s10620-017-4722-8. Epub 2017 Aug 23.

Early Colorectal Cancer Detected by Machine Learning Model Using Gender, Age, and Complete Blood Count Data.

Hornbrook MC[1], Goshen R[2], Choman E[2], O'Keeffe-Rosetti M[3], Kinar Y[2,4], Liles EG[3], Rust KC[3,5].

Author information

Erratum in

Correction to: Early Colorectal Cancer Detected by Machine Learning Model Using Gender, Age, and Complete Blood Count Data. [Dig Dis Sci. 2018]

Abstract

BACKGROUND: Machine learning tools identify patients with blood counts indicating greater likelihood of colorectal cancer and warranting colonoscopy referral.

AIMS: To validate a machine learning colorectal cancer detection model on a US community-based insured adult population.

METHODS: Eligible colorectal cancer cases (439 females, 461 males) with complete blood counts before diagnosis were identified from Kaiser Permanente Northwest Region's Tumor Registry. Control patients (n = 9108) were randomly selected from KPNW's population who had no cancers, received at ≥1 blood count, had continuous enrollment from 180 days prior to the blood count through 24 months after the count, and were aged 40-89. For each control, one blood count was randomly selected as the pseudo-colorectal cancer diagnosis date for matching to cases, and assigned a "calendar year" based on the count date. For each calendar year, 18 controls were randomly selected to match the general enrollment's 10-year age groups and lengths of continuous enrollment. Prediction performance was evaluated by area under the curve, specificity, and odds ratios.

RESULTS: Area under the receiver operating characteristics curve for detecting colorectal cancer was 0.80 ± 0.01. At 99% specificity, the odds ratio for association of a high-risk detection score with colorectal cancer was 34.7 (95% CI 28.9-40.4). The detection model had the highest accuracy in identifying right-sided colorectal cancers.

CONCLUSIONS: ColonFlag® identifies individuals with tenfold higher risk of undiagnosed colorectal cancer at curable stages (0/I/II), flags colorectal tumors 180-360 days prior to usual clinical diagnosis, and is more accurate at identifying right-sided (compared to left-sided) colorectal cancers.

Motivation: Colorectal Cancer is more treatable if detected earlier

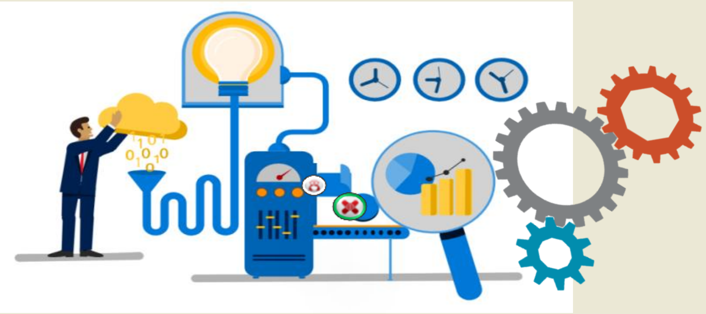
Colorectal cancer is the most commonest cancer in HK

5437 new cases of colorectal cancer in 2016

Screening / Examination:

Faecal occult blood

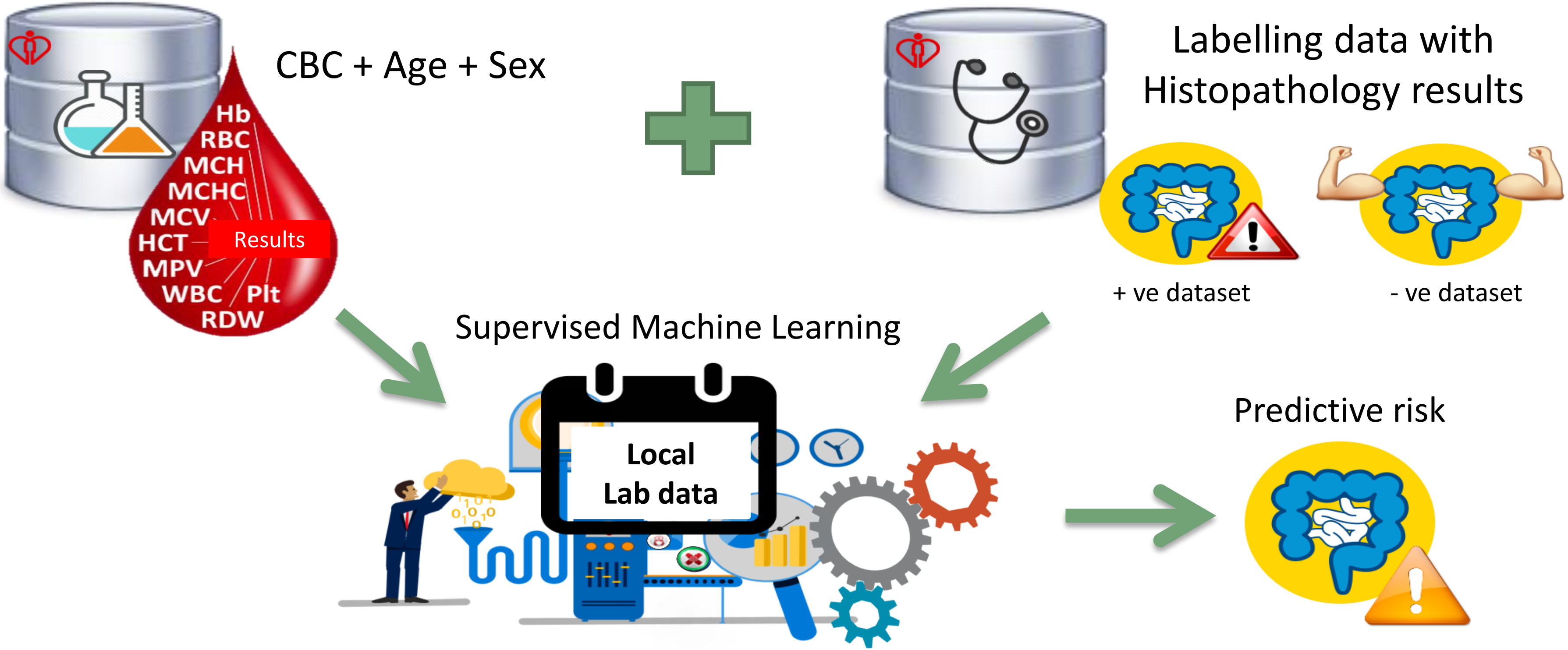
Colonoscopy



Can ML assist to find unscreened patients at high risk of colorectal cancer?

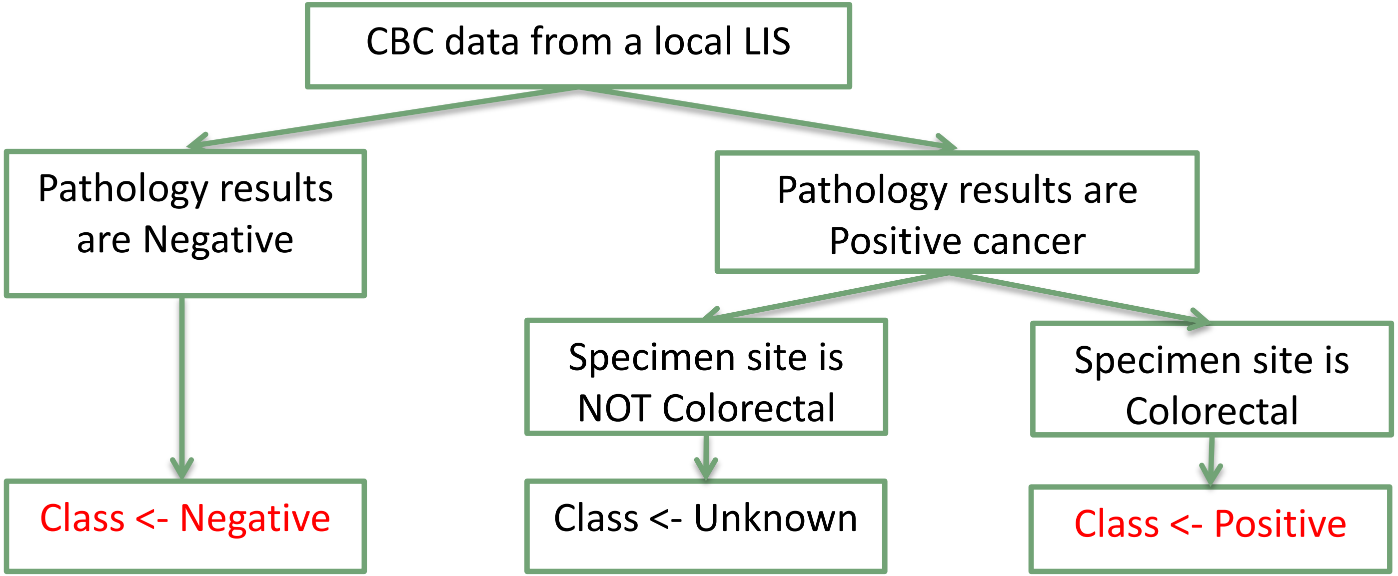
To recommend high risk patients to have a colonoscopy…

Training Dataset Preparation for Predictive Colorectal Cancer by Machine Learning



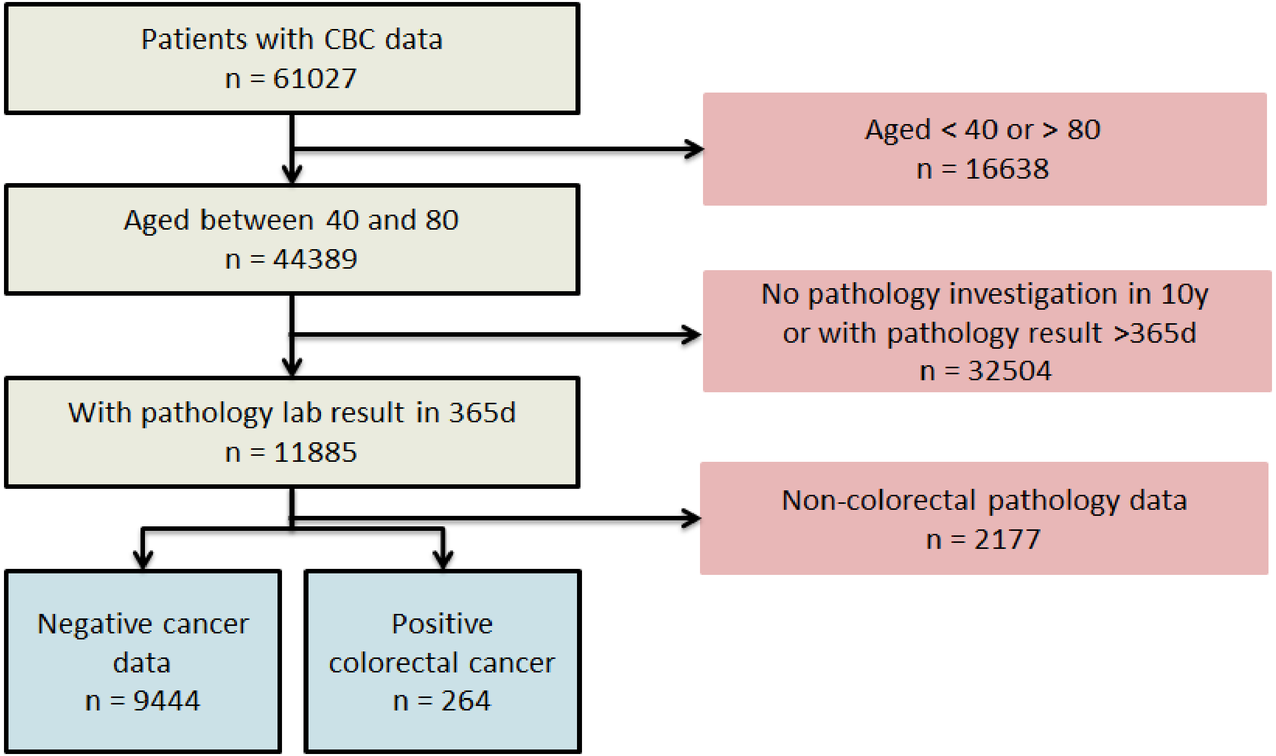
With ML algorithm, based on very subtle changes in CBC values to predict colorectal cancer

Data Extraction and Labelling



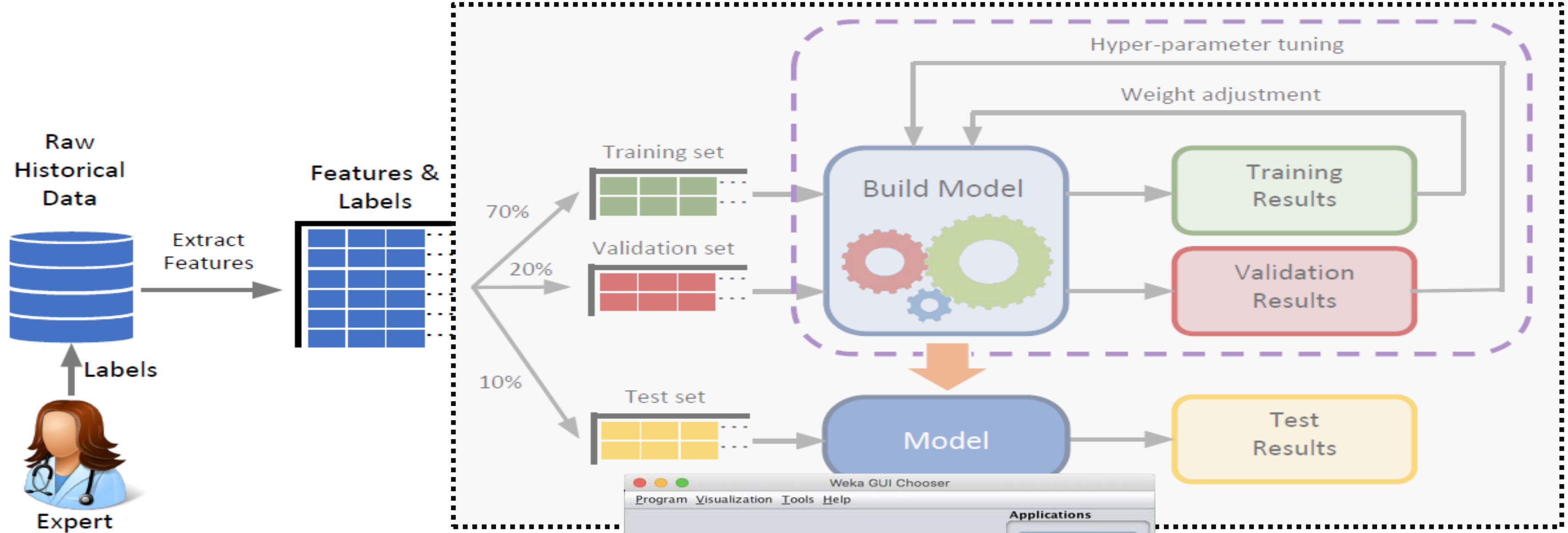
Training Dataset: De-identified lab data retrieved from Laboratory Information System of an acute hospital

Cohort Selection

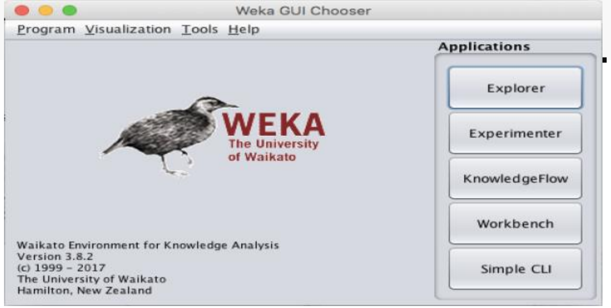


IT&HI DIVISION

Machine Learning Workflow



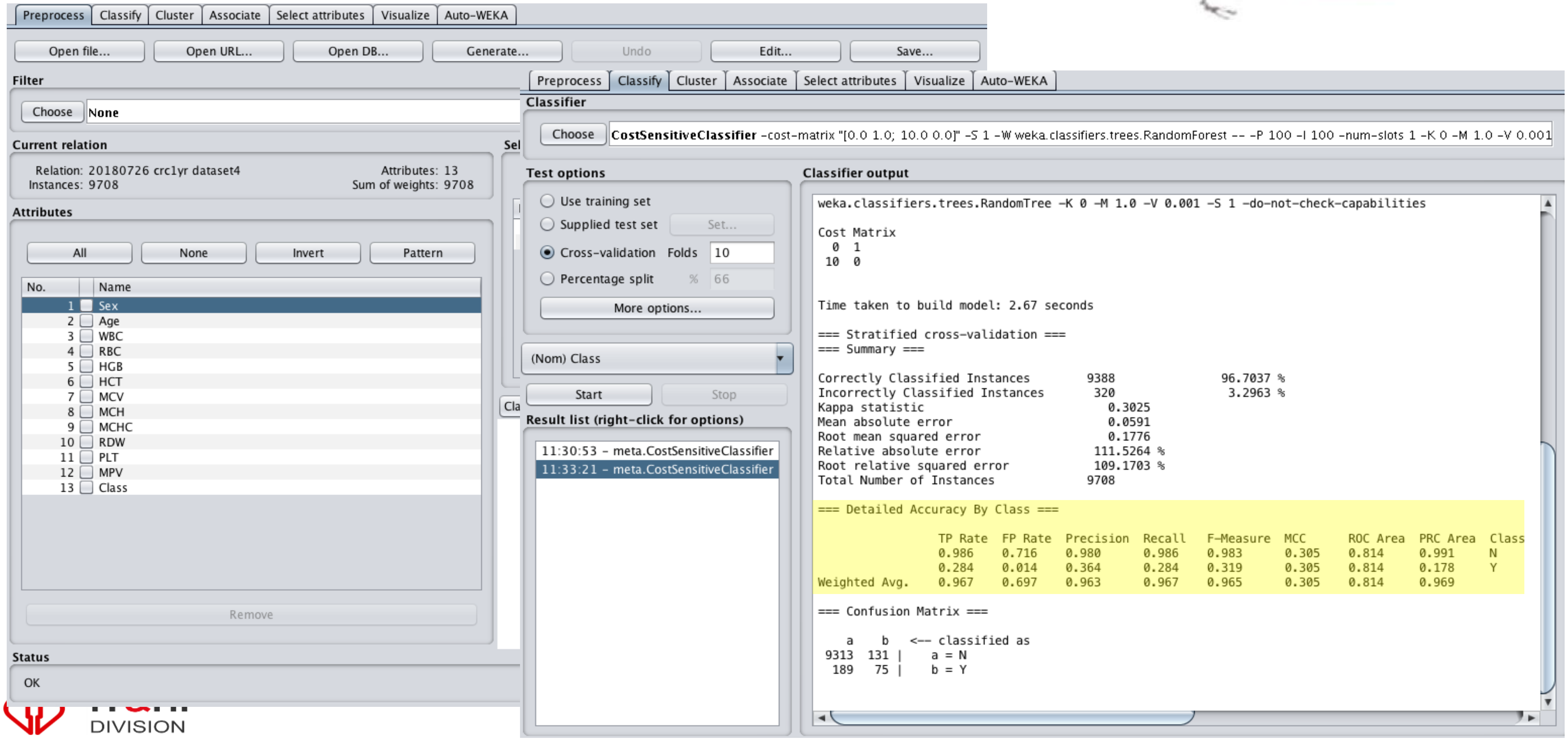
We tried using AutoML tools for the data modelling.



DataRobot

Figure adapted from https://www.capgemini.com/2016/05/machine-learning-has-transformed-many-aspects-of-our-everyday-life/

Data Modelling using



DIVISION

Evaluation Results from

Run Information	1.	2.	3.	4.
Scheme	Tree-J48	RandomForest	RandomForest	RandomForest +CostSensitiveClassifier (reweighted training)
Instances	9708 (Neg-9444; Pos-264)	9708 (Neg-9444; Pos-264)	9708 (Neg-9444; Pos-264)	9708 (Neg-9444; Pos-264)
Features	4 (Sex, Age, HGB, Class)	4 (Sex, Age, HGB, Class)	13 (Sex, Age, CBC, Class)	13 (Sex, Age, CBC, Class)
Test mode	10-fold CV	10-fold CV	10-fold CV	10-fold CV
Classification accuracy	97.84%	97.23%	96.67%	96.70%
TP Rate	N-1.000; P-0.208	N-0.994; P-0.216	N-0.987; P-0.235	N-0.986; P-0.284
FP Rate	N-0.792; P-0.000	N-0.784; P-0.006	N-0.765; P-0.013	N-0.716; P-0.014
Precision	N-0.978; P-1.000	N-0.978; P-0.483	N-0.979; P-0.339	N-0.980; P-0.362
Recall	N-1.000; P-0.208	N-0.994; P-0.216	N-0.987; P-0.235	N-0.986; P-0.284
F-Measure	N-0.989; P-0.345	N-0.986; P-0.298	N-0.983; P-0.277	N-0.983; P-0.319
AUC	0.581	0.685	0.781	0.814

Negative Predictive Value (NPV) – looks good

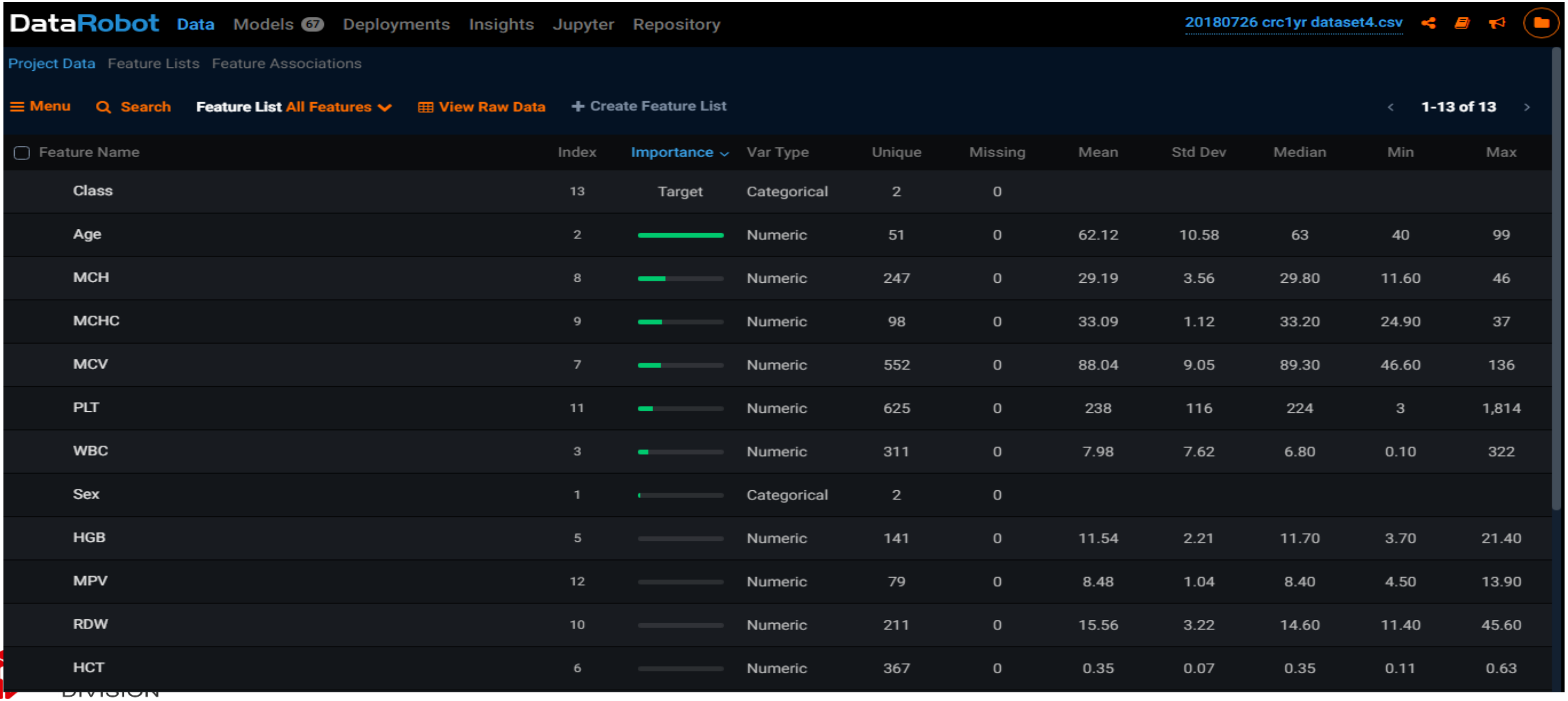
	Gold Standard			
		+	-	
Test	+	75	189	264
Result	-	132	9312	9444
		207	9501	9708

PPV= 0.284

NPV= 0.986

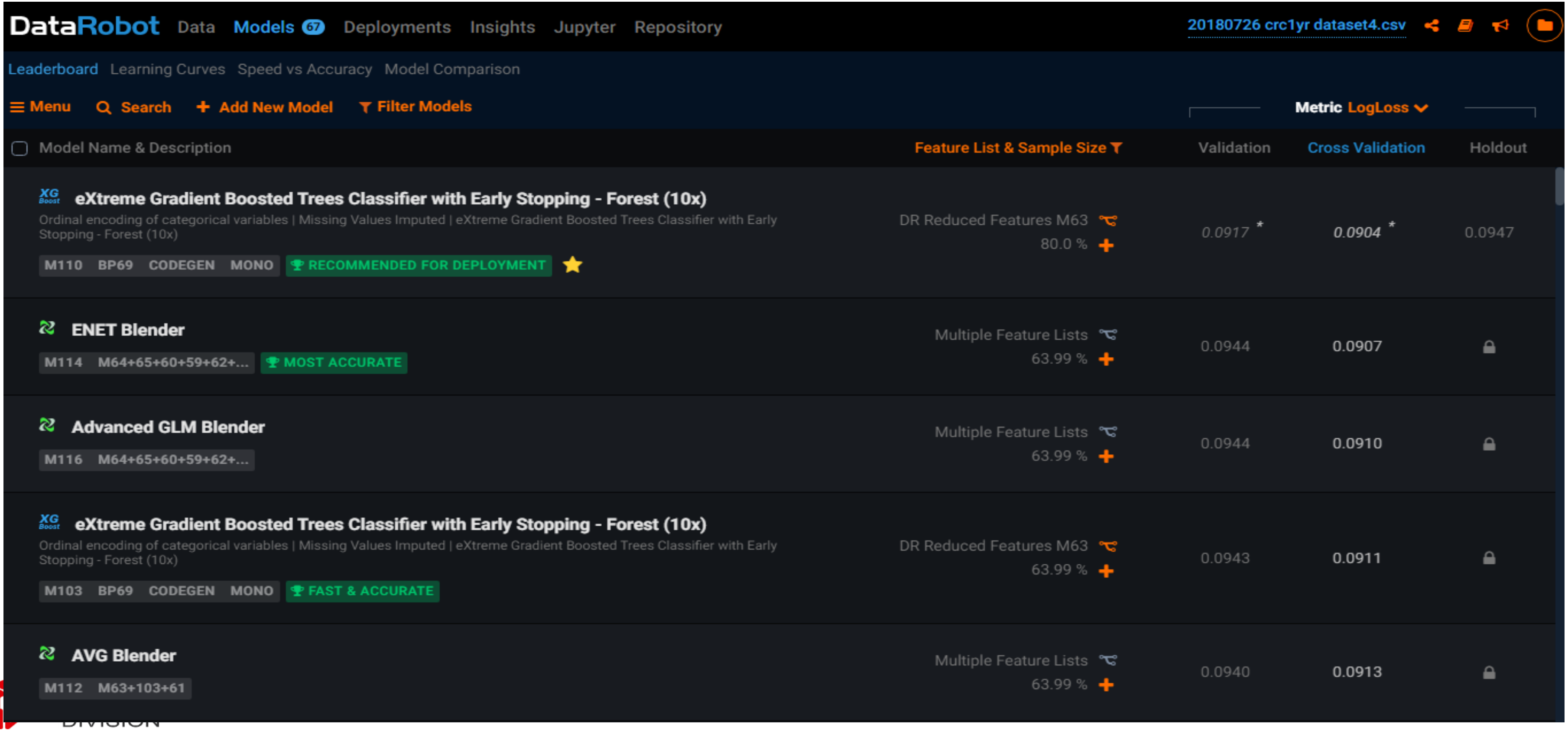
Sensitivity	Specificity
0.362	0.980

Rerun the dataset using DataRobot

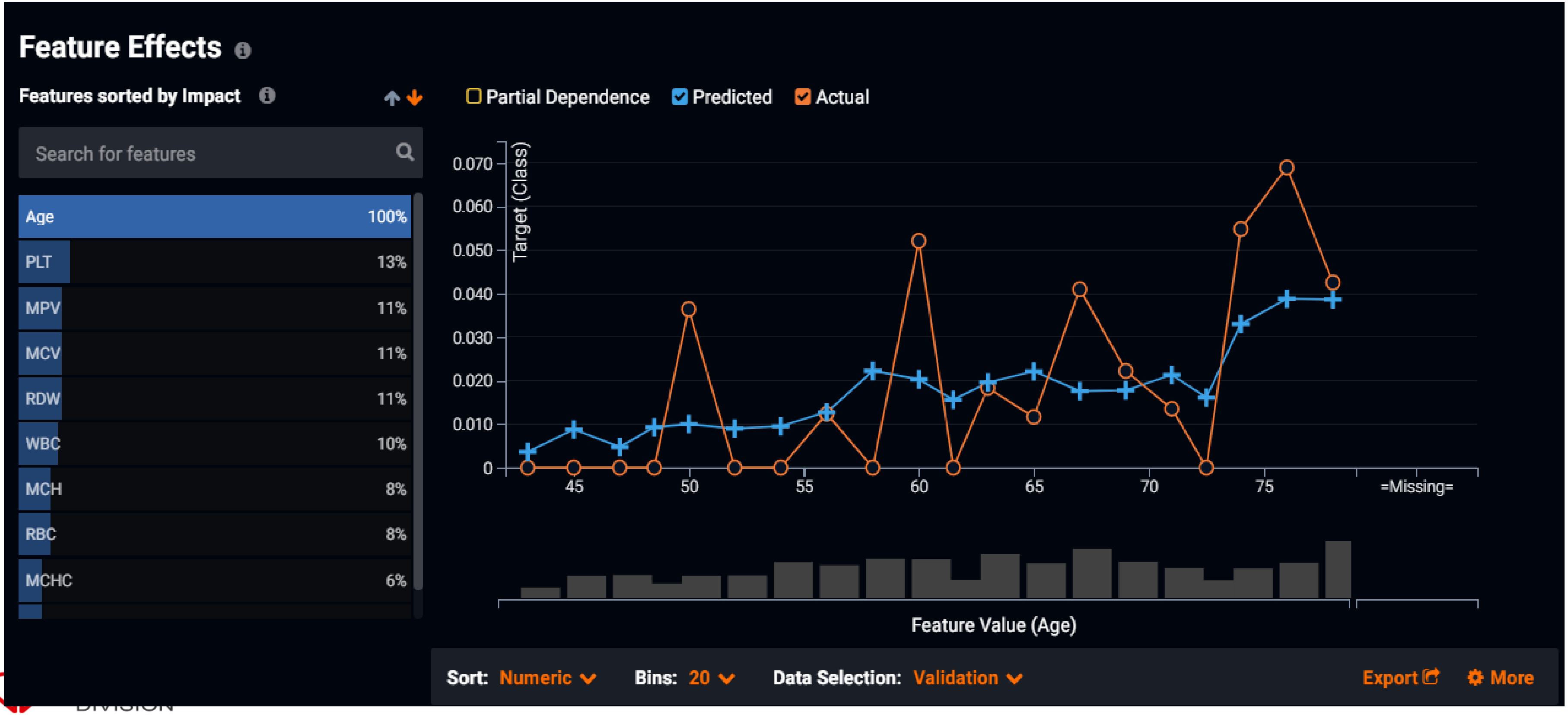


Feature Name	Index	Importance	Var Type	Unique	Missing	Mean	Std Dev	Median	Min	Max
Class	13	Target	Categorical	2	0					
Age	2		Numeric	51	0	62.12	10.58	63	40	99
MCH	8		Numeric	247	0	29.19	3.56	29.80	11.60	46
MCHC	9		Numeric	98	0	33.09	1.12	33.20	24.90	37
MCV	7		Numeric	552	0	88.04	9.05	89.30	46.60	136
PLT	11		Numeric	625	0	238	116	224	3	1,814
WBC	3		Numeric	311	0	7.98	7.62	6.80	0.10	322
Sex	1		Categorical	2	0					
HGB	5		Numeric	141	0	11.54	2.21	11.70	3.70	21.40
MPV	12		Numeric	79	0	8.48	1.04	8.40	4.50	13.90
RDW	10		Numeric	211	0	15.56	3.22	14.60	11.40	45.60
HCT	6		Numeric	367	0	0.35	0.07	0.35	0.11	0.63

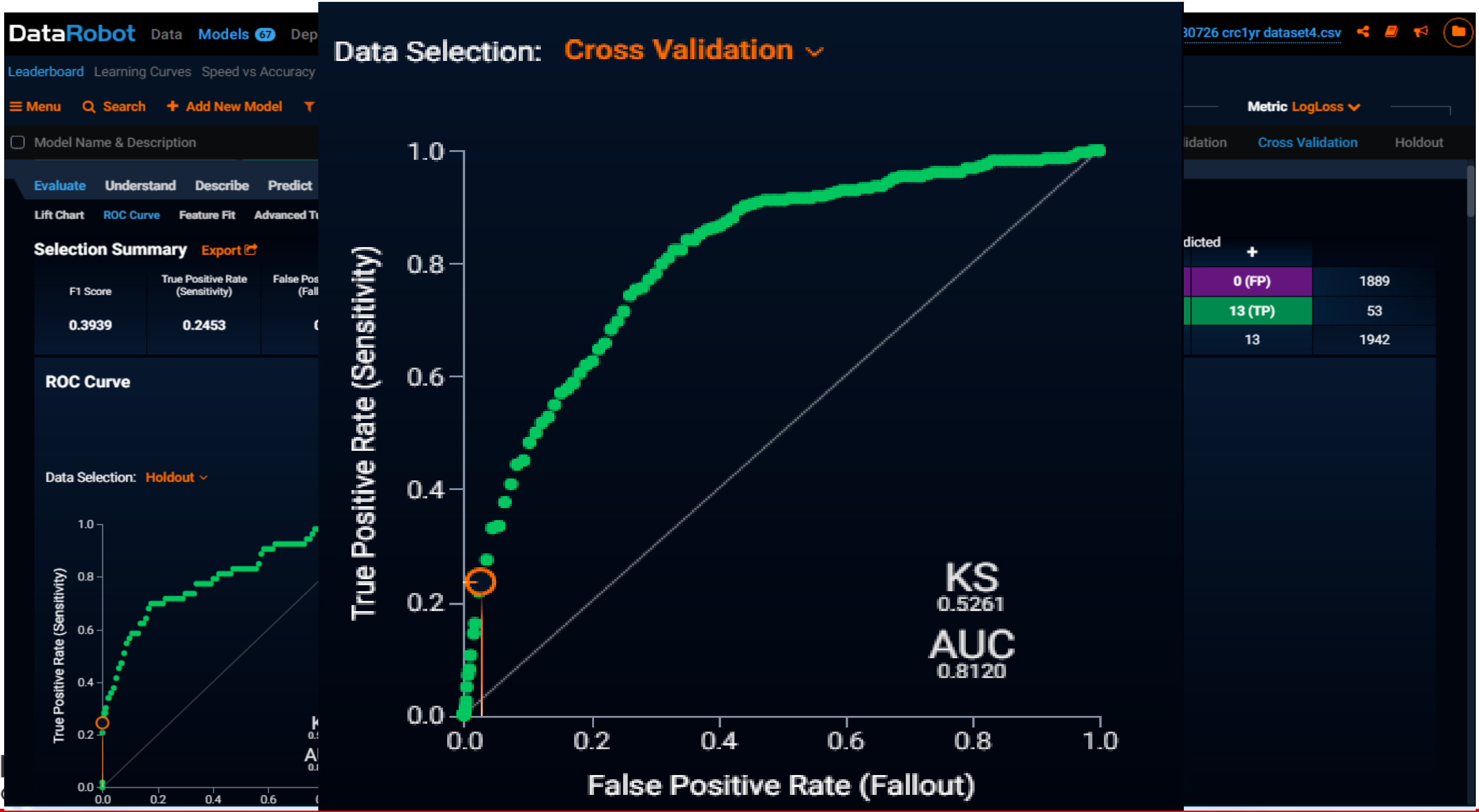
Automatic Data Modelling



Data Model – Feature Effects



Data Model Evaluation



IT&
DIVISI

Lessons learnt

- Importance of good quality data for Machine Learning
- Heavy work on data Retrieval and Labelling
- Features selection requires Domain Knowledge
- Validation is critically important
- Imbalanced dataset issue
- Easy-to-use Data Science tools available for data modelling
  → empowers ordinary people to take machine learning initiatives into their own hands

References

- *Hornbrook MC, Goshen R, Choman E, O'Keeffe-Rosetti M, Kinar Y, Liles EG, Rust KC.* Early Colorectal Cancer Detected by Machine Learning Model Using Gender, Age, and Complete Blood Count Data. Dig Dis Sci. 2017 Oct.
- *Kinar Y, Kalkstein N, Akiva P, Levin B, Half EE, Goldshtein I, Chodick G, Shalev V.* Development and validation of a predictive model for detection of colorectal cancer in primary care by analysis of complete blood counts: a binational retrospective study. J Am Med Inform Assoc. 2016 Sep; 23(5): 879–890.
- Weka. Waikato Environment for Knowledge Analysis https://www.cs.waikato.ac.nz/ml/weka/index.html
- *JEN UNDERWOOD*. White Paper: Moving from Business Intelligence to Machine Learning with Automation

IT&HI
DIVISION